

**Proceedings of the
17th International Conference on
Cyber Warfare and Security
State University of New York at Albany
Albany, New York, USA
17-18 March 2022**



**Dean Robert P. Griffin, Dr. Unal Tatar
and Dr. Benjamin Yankson**

A conference managed by ACI, UK

Proceedings of the

**17th International Conference on
Cyber Warfare and Security
ICCWS 2022**

**A Conference
Hosted By**

**State University of New York at Albany
Albany, New York
USA**

17-18 March 2022

Copyright the authors, 2022. All Rights Reserved.

No reproduction, copy or transmission may be made without written permission from the individual authors.

Review Process

Papers submitted to this conference have been double-blind peer reviewed before final acceptance to the conference. Initially, abstracts were reviewed for relevance and accessibility and successful authors were invited to submit full papers. Many thanks to the reviewers who helped ensure the quality of all the submissions.

Ethics and Publication Malpractice Policy

ACIL adheres to a strict ethics and publication malpractice policy for all publications – details of which can be found here:

<http://www.academic-conferences.org/policies/ethics-policy-for-publishing-in-the-conference-proceedings-of-academic-conferences-and-publishing-international-limited/>

Self-Archiving and Paper Repositories

We actively encourage authors of papers in ACIL conference proceedings and journals to upload their published papers to university repositories and research bodies such as ResearchGate and Academic.edu. Full reference to the original publication should be provided.

Conference Proceedings

The Conference Proceedings is a book published with an ISBN and ISSN. The proceedings have been submitted to a number of accreditation, citation and indexing bodies including ISI Web of Science and Elsevier Scopus. Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

The Electronic version of the Conference Proceedings is available to download from <https://papers.academic-conferences.org/>.

The Conference Proceedings for this year and previous years can be purchased from <http://academic-bookshop.com>

E-Book ISBN: 978-1-914587-27-6

E-Book ISSN: 2048-9889

Book version ISBN: 978-1-914587-26-9

Book Version ISSN: 2048-9870

Published by Academic Conferences International Limited

Reading, UK

+44118 324 6938

www.academic-conferences.org

info@academic-conferences.org

Contents

Paper Title	Author(s)	Page No
Preface		vii
Committee		ix
Biographies		xi
Research papers		
Cyberbullying Indicator as a Precursor to a Cyber Construct Development	Salam Khalifa Al-Romaihi and Richard Adeyemi Ikuesan	1
Education and Training Against Threat of Phishing Emails	Ladislav Burita, Ivo Klaban and Tomas Racil	7
Don't Drink the Cyber: Extrapolating the Possibilities of Oldsmar's Water Treatment Cyberattack	James Cervini, Aviel Rubin and Lanier Watkins	19
An Ontology for Effective Security Incident Management	Sabarathinam Chockalingam and Clara Maathuis	26
Bureau of Justice Student Computer and Digital Forensics Educational Opportunities Program: The Assessment of Online Graduate Students	Kyung-Shick Choi, Chitkushev Lou, Kyung-Seok Choo and Claire Seungeun Lee	36
Protecting Networks with Intelligent Diodes	Jason Dahlstrom and Stephen Taylor	45
Can Attrition Theory Provide Insight for Cyber Warfare?	Stephen Defibaugh and Donna Schaeffer	55
A Novel DevSecOps Model for Robust Security in an MQTT Internet of Things	Manasa Ekoramaradhya and Christina Thorpe	63
Exploring Ontologies for Mitigation Selection of Industrial Control System Vulnerabilities	Thomas Heverin, Michael Cordano, Andy Zeyher, Matthew Lashner and Sanjana Suresh	72
Future Needs of the Cybersecurity Workforce	Connie Justice, Char Sample, Sin Ming Loo, Alex Ball and Clay Hampton	81
Zero Trust and Advanced Persistent Threats: Who Will Win the War?	Bilge Karabacak and Todd Whittaker	92
Advancing Cybersecurity Capabilities for South African Organisations Through R&D	Zubeida Casmod Khan and Nenekazi Nokuthala Penelope Mkuzangwe	102
Zero Trust Container Architecture (ZTCA): A Framework for Applying Zero Trust Principals to Docker Containers	Darragh Leahy and Christina Thorpe	111
APT Cyber-Attack Modelling: Building a General Model	Martti Lehto	121
Design and Evaluation of a Cyber Protection Team Planner Work Aid	Kristen Liggett, Arielle Stephenson, Meghan Strang and Geoffrey Dobson	130
I Know You by Heart: Biometric Authentication based on Electrocardiogram (ECG) signals	Christoph Lipps, Lea Bergkemper, Jan Herbst and Hans Dieter Schotten	135
Shallow Deep Learning using Space-filling Curves for Malware Classification	David Long and Stephen O'Shaughnessy	145

Paper Title	Author(s)	Page No
Integrating Democratic Cybersecurity: Empowerment of Traditional Law Enforcement and Democratic Public Safety	Michael Losavio, Jeffrey C. Sun, Sharon Kerrick, Adel Elmaghraby, Cheryl Purdy and Clay Johnson	155
On Explainable AI Solutions for Targeting in Cyber Military Operations	Clara Maathuis	166
Analyzing the Performance of Block-Splitting in LLVM Fingerprinting	William Mahoney, Philip Sigillito, Jeff Smolinski, J. Todd McDonald and George Grispos	176
Taxonomy of Social Engineering Attacks: A Survey of Trends and Future Directions	Arianit Maraj and William Butler	185
Blockchain Technology for Addressing Privacy and Security Issues in Cloud Computing	Pardis Moslemzadeh Tehrani, Gabriele Kotsis and Andasmara Rizky Pranata	194
Context-Aware Cyber Threat Intelligence Exchange Platform	Michael Motlhabi, Phumeza Pantsi, Bokang Mangoale, Rofhiwa Netshiya and Samson Chishiri	201
The Development of Cybersecurity Awareness Measurement Model in the Water Sector	Bryan S. Mufor, Annlizé Marnewick and Suné von Solms	211
Circuit-Variant Moving Target Defense for Side-Channel Attacks	Tristen Mullins, Brandon Baggett, Todd R. Andel and J. Todd McDonald	219
A Modern ICT Network Simulator for Co-Simulations in Smart Grid Applications	Fabian Niehaus, Bastian Fraune, Giacomo Gritzan and Richard Sethmann	227
Towards Detection of Selfish Mining Using Machine Learning	Matthew Peterson, Todd Andel and Ryan Benton	237
Specialised Media Monitoring Tool to Observe Situational Awareness	Heloise Pieterse, Carien Van 't Wout, Zubeida Khan and Chris Serfontein	244
A New Dawn for Space Security	Jordan J. Plotnek and Jill Slay	253
Assessment of Cybersecurity Risks: Maritime Automated Piloting Process	Jouni Pöyhönen and Martti Lehto	262
Utilizing Switch Port Link State to Detect Rogue Switches	Travis Quitiquit and Vijay Bhuse	272
Aligning South African Data and Cloud Policy with the PoPI Act	Emma Raaff, Nicole Rothwell and Aidan Wynne	279
Transient Execution and Side Channel Analysis: A Vulnerability or a Science Experiment?	Michael Shepherd, Scott Brookes and Robert Denz	288
Bug Bounties: Between New Regulations and Geopolitical Dynamics	Jantje Silomon, Mischa Hansel and Fabiola Schwarz	298
Emerging Cyber risk Challenges in Maritime Transportation	Jussi Simola and Jouni Pöyhönen	306
Improving Protection Against Cybersecurity Attacks of Emergency Dispatch Centers	James Sweeney and Vu Tran	315
Increasing Industry Profitability and Cyber Hygiene Utilizing Awareness Progression Methods	John Thebarger, Mark Reith and Wayne Henry	325
Identifying Adversaries' Signatures Using Knowledge Representations of Cyberattack Techniques on Cloud Infrastructure	Gilliam van der Merwe, Christian Muller, Wilhelm van der Merwe and Dewald Blaauw	333

Paper Title	Author(s)	Page No
Rising Above Misinformation and Deepfakes	Namosha Veerasamy and Heloise Pieterse	340
Multi-Purpose Cyber Environment for Maritime Sector	Gabor Visky, Arturs Lavrenovs, Erwin Orye, Dan Heering, Kimberly Tam and Olaf M. Maennel	349
Ethical and Legal Aspects Pertaining to law Enforcement use of Drones	MM Watney	358
Social Media Privacy Using EDEE Security Model	Benjamin Yankson, Eric Cajigal Delgado, Amjad Al-Jabri, Natalie Gitin and Sydney Davidson	366
PHD Papers		
The Impact of CISO Appointment Announcements on the Market Value of Firms	Adrian Ford, Ameer Al-Nemrat, Seyed Ali Ghorashi and Julia Davidson	375
What is a Substantial Contribution to a Research Project in Offensive Cyberspace Operations that Merits Co-Authorship?	Gazmend Huskaj	385
Human Errors: A Cybersecurity Concern and the Weakest Link to Small Businesses	Tabisa Ncubekezi	395
Need for a Cyber Resilience Framework for Critical Space Infrastructure	Syed Shahzad, Li Qiao and Keith Joiner	404
Ransomware Detection using Process Memory	Avinash Singh, Richard Adeyemi Ikuesan and Hein Venter	413
Evaluating the Reliability of Android Userland Memory Forensics	Sneha Sudhakaran, Aisha Ali-Gombe, Andrew Case and Golden G Richard III	423
The Cumulative Cyber Deterrence	Maija Turunen and Martti J. Kari	433
Masters Research Papers		
Performance Implications for Multi-Core RISC-V Systems with Dedicated Security Hardware	Samuel Chadwick, Scott Graham and James Dean	440
Analysis of Image Thresholding Algorithms for Automated Machine Learning Training Data Generation	Tristan Creek and Barry E. Mullins	449
Securing InfiniBand Networks with the Bluefield-2 Data Processing Unit	Noah Diamond, Scott Graham and Gilbert Clark	459
Malware Binary Image Classification Using Convolutional Neural Networks	John Kiger, Shen-Shyang Ho and Vahid Heydari	469
Defending Small Satellites from Malicious Cybersecurity Threats	Banks Lin, Wayne Henry and Richard Dill	479
Improving Hardware Security on Talos II Architecture Through Boot Image Encryption	Calvin Muramoto, Stephen Dunlap and Scott Graham	489
Technical Analysis of Thanos Ransomware	Ikuomor Ogiriki, Christopher Beck and Vahid Heydari	497
Analysis of Sexual Abuse of Children Online and CAM Investigations in Finland	Johanna Parviainen and Jyri Rajamäki	505
Digital Risk Management: Investigating Human-Factor Security with a Behaviorist Approach	Ruan Pretorius and Dewald Blaauw	513

Paper Title	Author(s)	Page No
Ensuring the Security of Space Systems from Eavesdropping Attacks	Caleb Richardson, Mark Reith and Wayne Henry	522
Work In Progress Papers		
Mitigating Global Cyber Risk Through Bridging the National Incident Response Capacity Gap	Elisabeth Dubois and Unal Tatar	527
An Exploration on APTs in Biocybersecurity and Cyberbiosecurity	Xavier-Lewis Palmer, Lucas Potter and Saltuk Karahan	532
Late Submissions		
Cyberwarfare and its Effects on Critical Infrastructure	Humairaa Yacoob Bhaiyat and Siphesihle Philezwini Sithungu	536
Addressing the Skills Shortage in Cybersecurity	Gareth Davies, Angela Mison and Peter Eden	544
Critical Systems Protection (CSP): The US Secret Service's Tactical Cyber Capability for Securing Protectees	Austin Hyman, Brian Nussbaum, Mario Bencivenga and Zachary Rizzo	552
WFH, not WTH? The Security Challenges of Working-From-Home	Neal Kushwaha, Piret Pernik and Bruce W. Watson	559
The Failure of Trust in Trusted Systems	Angela Mison, Gareth Davies and Peter Eden	568
New Wave Cyber Attacks	Angela Mison, Gareth Davies and Peter Eden	576
The Role of Big Tech in Future Cyber Defence	Angela Mison, Gareth Davies and Peter Eden	583
Data Mining for the Security of Cyber Physical Systems Using Deep-Learning Methods	Bhagawan Nath, Timo Hamalainen and Soundararajan Ezekiel	591
'Out Beyond Jointery': Developing a Model for Gaming Multi-Domain Warfare	Keith Scott	599
Research Gaps and Opportunities for Secure Access Service Edge	Stephanus Petrus van der Walt and Hein Venter	609
Abstracts Only		
OSINT on the Dark Web: Child Abuse Material Investigations	Iiro Lahti and Jyri Rajamäki	
Incremental Malware Detection and Classification Using Hidden Markov Models	Ran Liu and Charles Nicholas	
Cyber Warfare, An Existential Crisis	Jacob Oakley	
Benefits of Crowdsourced Penetration Testing: Implementing an Industry Wide Crowdsourced Platform to Strengthen Penetration Testing and Artificial Intelligence	Bianca Pasikhani	
Militarization of Police, Violence Against Blacks Counterinsurgency and Information Warfare: Ethical and Anticipated Ethical Issues	Richard Wilson and Michael Shifflett	
The Big Lie, Disinformation, and Insurgency: Ethical and Anticipatory Ethical Issues	Richard Wilson	

Paper Title	Author(s)	Page No
ISIS and U. S. Domestic Terrorism: A Comparison and Anticipating the Ethical Issues	Richard Wilson	
KKK, Hegel and The Alt Right: An Ethical and Anticipated Ethical Analysis	Richard Wilson	
Oath Keepers, Domestic Terrorism and Insurgency: Ethical and Anticipated Ethical Issues	Richard Wilson and Michael Shifflett	
Additional Materials		
Participant List		
Google Scholar	The Importance of Paper citations and Google Scholar	
About ACI		
Research Jotter		

ICCWS Preface

These proceedings represent the work of contributors to the 17th International Conference on Cyber Warfare and Security (ICCWS 2022), hosted State University of New York at Albany, USA on 17-18 March 2022. The Conference Chair is Dean Robert P. Griffin, College of Emergency Preparedness, Homeland Security, and Cyber Security (CEHC), and the Program Co-Chairs are Dr. Unal Tatar, from University at Albany and Dr. Benjamin Yankson, from College of Emergency Preparedness Homeland Security and Cybersecurity

ICCWS is a well-established event on the academic research calendar and now in its 17th year, the key aim remains the opportunity for participants to share ideas and meet the people who hold them. The scope of papers will ensure an interesting two days. The subjects covered illustrate the wide range of topics that fall into this important and ever-growing area of research.

The opening keynote presentation is given by Heather Hage, Griffiss Institute (GI), USA. There will be an address later in the morning by Tony Sager, Senior Vice President & Chief Evangelist, Center for Internet Security, USA on the topic of *Stranger in a Changed Land*. The second day of the conference will open with an address by Col. Jeffrey Erickson, the director of the Army Cyber Institute. His talk is entitled *Developing the Critical Capabilities Needed to Respond to Cyber Attacks on Modern Cities*.

With an initial submission of 120 abstracts, after the double blind, peer review process there are 52 Academic research papers, 7 PhD research papers, 10 Masters Research papers and 2 work-in-progress papers published in these Conference Proceedings. These papers represent research from Australia, Canada, Czech Republic, Estonia, Finland, Germany, Ireland, Kosovo, Malaysia, Netherlands, Norway, Pakistan, Qatar, South Africa, Sweden, the UK and the USA

We hope you enjoy the conference.

Dean Robert P. Griffin

State University of New York at Albany
Albany, New York, USA
March 2022

ICCWS Conference Committee

Prof Azween Abdullah, Taylors University, Malaysia; Dr Bulent Acma, Anadolu University, Eskisehir, Turkey; Dr William ("Joe") Adams, Univ of Michigan/Merit Network, USA; Prof Hamid Alasadi, Iraq University college, Iraq; Dr Elie Alhajjar, USMA, USA; Prof Todd Anel, University of South Alabama, USA; Prof Antonios Andreatos, Hellenic Air Force Academy, Greece; Leigh Armistead, Oak Ridge National Laboratory, USA; Researcher Jawad Awan, Institute of Information & Communication Technology, Pakistan; Prof Richard Baskerville, Georgia State University, USA; Dr Zakariya Belkhamza, Ahmed Bin Mohammed Military College, Qatar; Prof Vijay Bhuse, Grand Valley State University, USA; Dr Raymond Buettner, Naval Postgraduate School, USA; Dr Mustafa Canan, Naval Postgraduate School, USA; Dr Jim Chen, U.S. National Defense University, USA; Dr Nathan Clarke, University of Plymouth, UK; Dr Ronen Cohen, Ariel University Centre, Israel; Mr Edwin Covert, WarnerMedia, USA; Dr Paul Crocker, University of Beira Interior, Portugal; Dr Michael Dahan, Sapir College, Israel; Evan Dembskey, UNISA, South Africa; Prokopios Drogkaris, University of the Aegean, Greece; Prof Mariki Eloff, University of South Africa, South Africa; Prof Eric Filiol, ENSIBS, Vannes, France & CNAM, Paris, France; Larry Fleurantin, Fleurantin, Francois & Antonin, P.A., North Miami Beach, USA; Dr Ahmad Ghafarian, University of North Georgia, USA; Dr Scott Graham, Air Force Institute of Technology, USA; Prof Dr Tim Grant, Retired But Active Researcher, The Netherlands; Dr John Gray, Nova Southeastern University, USA; Virginia Greiman, Boston University, USA; Dr Michael Grimaila, Air Force Institute of Technology, USA; Dr Per Gustavsson, Combitech / Swedish Defence University / George Mason Univeristy, Sweden; Dr Ulrike Hugl, University of Innsbruck, Austria; Prof Bill Hutchinson, Edith Cowan University, Australia; Dr Berg Hyacinthe, State University of Haiti, Haiti; Ramkumar Jaganathan, VLB Janakiammal College of Arts and Science (affiliated to Bharathiar University), India; Dr Ezhil Kalaimannan, University of West Florida, USA; Dr Bilge Karabacak, Franklin University, USA; Dr Saltuk Karahan, Old Dominion University, USA; Prof Jesuk Ko, Universidad Mayor de San Andres, Bolivia; Dr Anne Kohnke, Lawrence Technological University, USA; Dr Ahmet Koltuksuz, Yasar University, Turkey; Dr Maximiliano Korstanje, University of Palermo, Buenos Aires, Argentina, Argentina; Prashant Krishnamurthy, University of Pittsburgh, USA; Mr Peter Kunz, DoctorBox, Germany; Rauno Kuusisto, Finnish Defence Force, Finland; Dr Gregory Laidlaw, University of Detroit Mercy, USA; Dr Arash Lashkari, UNB, Canada; Dr Sylvain (Sly) Leblanc, Royal Military College of Canada, Canada; Louise Leenen, CSIR, Pretoria, South Africa; Prof Martti Lehto, University of Jyväskylä, Finland; Dr Antoine Lemay, CYDEF, Canada; Dr Andrew Liaropoulos, University of Piraeus, Greece; Mr Trupil Limbasiya, NIIT University, Neemrana, Rajasthan, India; Christoph Lipps, German Research Center for Artificial Intelligence, Germany; Volodymyr Lysenko, University of Washington, USA; Dr Bill Mahoney, University of Nebraska, Omaha, USA; Dr Naufal Mansor, Kampus Uniciti Alam,, Malaysia; Dr Naufal Mansor, Kampus Uniciti Alam,, Malaysia; ASSI Haribabu Maruturi, qiscet, india; Dr Paul Maxwell, Army Cyber Institute, USA; Dr Todd McDonald, Air Force Institute of Technology, USA; Dr Robert Mills, Air Force Institute of Technology, USA; Dr Pardis Moslemzadeh tehrani, University of Malaya, malaysia; Dr Barry Mullins, Air Force Institute of Technology, USA; Prof Antonio Muñoz, University of Málaga, Spain; Dr Lilian Nassif, Public Ministry of Minas Gerais, Brazil; Dr Asoke Nath, St. Xavier's College(Autonomous), India; Dr Funminiye Olajide, Nottingham Trent University, UK; Mr Arif Mohamed Ismail Oliullah, Jefferies International Lts, UK; Prof Abdelnaser Omran, School of Economics, Finance and Banking, Universiti Utara Malaysia, Malaysia; Prof Dr Frank Ortmeier, Otto-von-Guericke Universität, Magdeburg, Germany; Rain Ottis , Tallinn University of Technology, Estonia; Prof Evgeny Pashentsev,

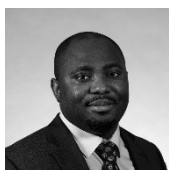
Diplomatic Academy of the Ministry of Foreign Affairs of the Russian Federation, Russia; Dr Gilbert Peterson, Air Force Institute of Technology, Dayton, Ohio, USA; Andy Pettigrew, George Washington University, USA; Dr Jackie Phahlamohlaka, Council for Scientific and Industrial Research, Pretoria, South Africa; Ms Heloise Pieterse, CSIR, South Africa; Dr Ajeet Singh Poonia, Bikaner Technical University, India; Dr Bernardi Pranggono, Sheffield Hallam University, UK; Prof Carlos Rabadão, Politechnic of Leiria, Portugal; Dr Trishana Ramluckan, University of KwaZulu-Natal, South Africa; Dr Mark Reith, Air Force Institute of Technology, USA; Lt Col Ernest Robinson, U.S. Marine Corps / Air War College, USA; Dr Neil Rowe, US Naval Postgraduate School, Monterey, USA; Prof Lili Saghafi, Canadian International College, Montreal, Canada; Dr Char Sample, US Army Research Laboratory, USA; Prof Leonel Santos, Politechnic of Leiria, Portugal; Ramanamurthy Saripalli, Pragati Engineering College, India; Dr Mark Scanlon, University College Dublin, Ireland; Dr Yilun Shang, Singapore University of Technology and Design, Singapore; Dr Dan Shoemaker, Centre for Assurance Studies, USA; Mr Paul Simon, Air Force Institute of Technology, USA; Mr Avinash Singh, University of Pretoria, South Africa; Dr Elena Sitnikova, University of South Australia, Australia; Prof Jill Slay, University of South Australia, Australia; Ass Prof Dr Risby Sohaimi, National Defence University of Malaysia, Malaysia; Dr William Spring, University of Hertfordshire, UK; Dr Kevin Streff, Dakota State University, USA; Ms sneha sudhakaran, Louisiana State University, USA; Dr Arwin Sumari, State Polytechnic of Malang, Java, Indonesia; Dr James Sweeney, Capella University, USA; Mr Unal Tatar, University at Albany - SUNY, USA; Pardis Moslemzadeh Tehrani, University of Malaya,; Peter Thermos, Columbia University/Palindrome Technologies, USA; Dr Bhavani Thuraisingham, University of Texas at Dallas, USA; Mr Patrick Tobin, University College Dublin, Ireland; Dr Antonio Jorge Tomeu-Hardasmal, University of Cadiz, Spain; Dr Vu Tran, Naval Air Warfare Center - Weapons Division, USA; Dr Eric Trias, Air Force Institute of Technology, USA; Dr Chia-Wen Tsai, Ming Chuan University, Taiwan; Brett van Niekerk, Durban University of Technology., South Africa; Dr Namosha Veerasamy, Council for Scientific and Industrial Research, South Africa; Stylianos Vidalis, School of Computer Science, University of Hertfordshire, UK; Prof Murdoch Watney, University of Johannesburg, South Africa; Richard Wilson, Towson University, USA; Hongyi Wu, Old Dominion University, Virginia, USA; Enes Yurtoglu, Turkish Air War College, Turkey; Dr Zehai Zhou, University of Houston-Downtown, USA

Biographies

Conference and Programme Chairs



Dr. Unal Tatar is currently an Assistant Professor of Cybersecurity at the College of Emergency Preparedness, Homeland Security, and Cybersecurity, University at Albany. He has 16+ years of cybersecurity experience of cybersecurity in government, industry, and academia. He is the former coordinator of the National Computer Emergency Response Team of Turkey. Dr. Tatar's research is funded by NSF, NSA, DOD, and Society of Actuaries. Dr. Tatar holds a BSc degree in Computer Engineering, an MS degree in Cryptography, and a Ph.D. in Engineering Management and Systems Engineering. His main topics of interest are cybersecurity risk management, cyber resiliency, cyber insurance, and blockchain.

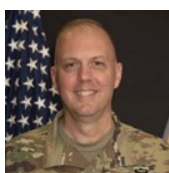


Dr. Benjamin Yankson is an Assistant Professor of Cybersecurity at the College of Emergency Preparedness Homeland Security and Cybersecurity. He has over 15yrs experience in various technical leadership roles in Information Technology security within Healthcare and Education. He is the former Application Manager, Critical Care Information System for the province of Ontario's (CitiCall Ontario), Canada. Dr. Yankson holds a CompTIA Security+, a BA degree in Information Technology, a master's degree in Information Technology Security (MITS), and a Ph.D. Computer Science. His current teaching and research work focuses on IoT Security, Cybersecurity Risk Management, Threat Risk Assessment (TRA), Security Auditing/Compliance, Digital Forensics, and Privacy.



Robert Griffin currently serves as the Founding Dean of the College of Emergency Preparedness, Homeland Security, and Cyber Security (CEHC). He came to CEHC after a long career in homeland security at the Federal and local levels of government. In the Federal Government, Dr. Griffin served as the Under Secretary (Acting) for Science and Technology at the US Department of Homeland Security (DHS), the Deputy Under Secretary for Science and Technology, and the Director of the Science and Technology Directorate's First Responders Group. While at Science and Technology, he managed a diverse \$1.1 billion dollar portfolio of homeland security research. Dr. Griffin was recognized for his exceptional management and leadership skills in 2016 when President Obama awarded him the Presidential Rank Award (Distinguished), the Nation's highest award for career members of the Senior Executive Service.

Keynote Speakers



Colonel Jeffrey M. Erickson is the Director of the Army Cyber Institute at the United States Military Academy (USMA) located at West Point, New York. In his position as Director, COL Erickson leads a 50-person, multi-disciplinary research institute focused on expanding the Army's knowledge of the cyberspace domain. He began his Army career as an Armor officer before transitioning to the Simulation Operations functional area, where for the last 15 years he has been using simulations to train from the individual up to the Joint and Combatant Command levels. He has a B.S. in Computer Science from the United States Military Academy, an M.S. in Management Information Systems from Bowie State University, and a M.S. in National Resource Strategy from the Eisenhower School (formerly the Industrial College of the Armed Forces). His fields of interest are simulations for live-virtual-constructive training, testing, and wargaming.



Heather Hage is President & CEO of the Griffiss Institute (GI), a leading national nonprofit talent and technology accelerator for the United States Department of Defense. Heather leads the expansion of the GI's portfolio of international talent and technology development programs, and the standup of Air Force Research Laboratory Information Directorate's new "over the fence" R&D and incubator facilities at Innovare Advancement Center in Rome, NY. Prior to joining the Griffiss Institute, Heather served the State University of New York for 17 years, where she oversaw innovation and partnerships for the SUNY System. A graduate of Hamilton College and Albany Law School, where she was Managing Editor of the Journal of Science and Technology, Heather's professional background includes media relations, telecommunications, IP management and corporate transactions. An adjunct faculty member at Albany Law School and SUNY Polytechnic Institute's College of Nanoscale Science and Engineering, Heather is the recipient of various honors and awards, including SUNY's 2015 Woman of Excellence award for outstanding service, exemplary support of women in leadership, and a distinguished career.



Tony Sager is a Senior Vice President and Chief Evangelist for CIS® (The Center for Internet Security, Inc.). He leads the development of the CIS Controls™, a worldwide consensus project to find and support technical best practices in cybersecurity. Sager champions use of CIS Controls and other solutions gleaned from previous cyber-attacks to improve global cyber defense. He also nurtures CIS' independent worldwide community of volunteers, encouraging them to make their enterprise, and the connected world, a safer place. In addition to his duties for CIS, he is an active volunteer in numerous community service activities: the Board of Directors for the Cybercrime Support Network; and a member of the National Academy of Sciences Cyber Resilience Forum; DHS Cyber Safety Review Board. Sager retired from the National Security Agency (NSA) after 34 years as an Information Assurance professional. Sager's awards and commendations at NSA include the Presidential Rank Award at the Meritorious Level, twice, and the NSA Exceptional Civilian Service Award.

Mini Track Chairs



Dr. Jawad Hussain Awan is a member of IFIP WG 9.10: ICT Uses in Peace and War, and an active researcher with the Global Foundation for Cyber Studies and Research Group. His research interests are ICT, Data Science, Cyber security, Security challenges in Smart and Physical Systems. He has published his research in several national and international research journals. He is the member of Editorial Review Board of a number of journals including of IEEE Access, Egyptian Journal of Informatics and the Journal of Information Warfare. He performed his duties as Committee member in ICCWS 2019, 2020, 2021 and 2022. He is also Microsoft Certified Professional in Web Programming, and has won research awards with the collaboration of HEC, Pakistan and the British Council. He also worked as Lab Manager, Lecturer, and Research Fellow.



Dr Elena Sitnikova is an academic at UNSW Canberra. She is a global research leader in Critical Infrastructure protection currently focusing on intrusion detection for SCADA systems cybersecurity, cyber-physical systems and industrial IoT(IIoT). Her research explores systems resilience and the use of artificial intelligence in the analytics of abnormal data that may affect critical systems. She is one of the first Australians to be certified in Certified Secure Software Lifecycle Professional (CSSLP). Elena is holding a Senior Fellowship of the Higher Education Academy(SFHEA) and Australian Office for Learning and Teaching(OLT) Team Citation award for Outstanding Contributions to Student Learning.



Dr. Pardis Moslemzadeh Tehrani, is a Senior Lecturer at the Faculty of Law, University of Malaya where she has been since 2015. Her research interests span in the area of Information Technology Law and International Humanitarian Law. Pardis has made scholarly contributions in peer-reviewed Journals and presented papers in a number of conferences and served on many program committees. She is an International

Scientific Member of the Australian and New Zealand Society of International Law and currently a member of Editorial Review Board of International Journal of Digital Crime and Forensics.

Workshop Facilitator



Dr Edwin “Leigh” Armistead is the Principal of ArmisteadTec LLC, a Veteran-Owned Small Business (VOSB) that is certified by the Small Business Administration (SBA) as a Historically Underutilized Business Zone (HBZ), <https://www.armisteadtec.com/>. This company primarily focuses on the academic functions of Information Warfare (IW), to include curriculum development, teaching as well as other related tasks to include

lecturing and writing. A retired United States Naval Officer, Dr Armistead has significant Information Operations academic credentials having written his PhD on the conduct of Cyber Warfare by the federal government and has published three books, in an unclassified format in 2004, 2007 and 2010, all focusing on full Information Warfare. He is also the Chief Editor of the Journal of Information Warfare (JIW) <https://www.jinfowar.com/>; the Program Director of the International Conference of Cyber Warfare and Security and the Vice-Chair Working Group 9.10, ICT Uses in Peace and War.

Biographies of Contributing Authors

Vijay Bhuse is an Associate Professor of Computer Science at the Grand Valley State University. He received his Ph.D. from Western Michigan University in 2007 and completed his postdoctoral fellowship from the Dartmouth College. He worked in industry before returning to academia. His research interests are Network Security, Wireless Sensor Networks and Secure Coding.

Dewald Blaauw I currently hold a full-time lecturer position at the University of Stellenbosch and my primary research focus is Cyber Security, E learning, Electronic business/commerce, Computer Security Engineering, Cyber forensics, Information Technology Project management, Information systems management and System Design.

Scott Brookes is a Principal Scientist in the Trusted and Resilient Systems group at Riverside Research. He earned his Ph.D. from Dartmouth College in 2018, where he studied secure operating system and hypervisor architectures. Dr. Brookes has numerous first and second author publications in the areas of secure operating systems and hypervisors.

Ladislav Burita, University of Defence, Brno, Czech Republic. PhD. from 1985 and professor from 2003. He worked in NATO/MIP team; was head of the faculty research program and defense project MENTAL. His area of interest are information and knowledge systems. He published hundred papers at conferences and in journals.

James Cervini is a Doctor of Engineering student at Johns Hopkins University. His research interests include cyber-physical system security, fog-computing, virtualization, and penetration testing. James received his M.S. from JHU’s Information Security Institute where he completed his capstone project involving a security analysis of wireless smart-meter systems.

Samuel D. Chadwick is a graduate scholar pursuing a master’s degree in Computer Engineering at the Air Force Institute of Technology, USA. He earned his B.S. in Computer Engineering from the United

States Air Force Academy in 2018. His research interests include secure processor architecture, computer networks, and cyber security.

Sabarathinam Chockalingam is a Research Scientist at the Institute for Energy Technology in Halden, Norway. Saba has a PhD in Cyber Security from the Delft University of Technology and MSc in Cyber Security and Management from the University of Warwick. His research interests include cyber security, risk management and safety.

Kyung-Shick Choi, PhD is a Professor of the Practice and the Program Director at Boston University, where he oversees the graduate programs in Cybercrime Investigation and Cybersecurity. He established a track record in designing and delivering law enforcement training programs in cybercrime investigations. He is currently working on federally funded cybercrime- and cybersecurity-related projects.

Mike Cordano is a Cybersecurity and Mechanical Engineer at Gryphon Technologies, a defense company. He specializes in providing security for U.S. Navy control-systems. He holds a BS in Mechanical Engineering (2018) from Rowan University and is currently writing a capstone thesis for his MS in Cybersecurity, concentrating on Cyber Terrorism.

Tristan Creek is completing his Masters degree in Cyber Operations at the Air Force Institute of Technology, Wright-Patterson AFB, OH, USA. Research interests include signal intelligence, security posturing, and network security.

Gareth Davies is the Director of the Cybercrime and Digital Forensic laboratories at the University of South Wales. As Chairman of The First Forensic Forum (F3), Gareth has organised international conferences and regular yearly workshops to demonstrate new digital forensics research and cutting edge technologies to the law enforcement community.

Stephen Defibaugh currently holds a Doctorate in Science (DS.c) in Cyber Security from Marymount University and is an active member in Phi Kappa Phi. He has focused his research in the realm of cyber warfare and cyber maneuvering. He is a currently a Deputy CISO for the Federal Government and a Marymount University Research Fellow.

Robert Denz is the director of the Trusted and Resilient Systems research at Riverside Research. In this role, he leads a team of researchers that ensure the provenance, security, reliability, and resilience of military systems. He received his PhD. from Thayer School of Engineering at Dartmouth College in secure hypervisor and kernel design (2016).

Noah Diamond is a graduate student pursuing a master's degree in Computer Engineering at the Air Force Institute of Technology, USA. Mr. Diamond has a B.S. in Computer Engineering from the United States Air Force Academy. His research focus includes the InfiniBand Architecture, computer networks, and critical infrastructure security.

Richard Dill, USAF. Major Dill is an Assistant Professor of Computer Science at the Air Force Institute of Technology, Wright-Patterson AFB, Ohio USA. He received his B.S.C.S. from the University of Maryland at College Park in Maryland, and a Master's and Ph.D. in Computer Science from AFIT. His research interests include computer security, algorithms, wargaming, and artificial intelligence.

Adrian Ford is an information technology manager with over 25 years' experience and a doctoral research student of information security at the University of East London. He holds an MBA from Lancaster University Management School (2009), professional membership of the British Computer Society (MBCS) and is a Freeman of the Worshipful Company of Information Technologists.

Bastian Fraune holds a M.Sc. in Computer Science and works as a scientific researcher at the City University of Applied Sciences in Bremen. His research interests include cybersecurity in general, as well as trusted computing based concepts. His ongoing doctoral thesis deals with research questions on IT security monitoring.

George Grispos Assistant Professor of Cybersecurity in the School of Interdisciplinary Informatics (University of Nebraska, Omaha). Received a MSc. in Computer Forensics and E-Discovery and his Ph.D. in Computing Science (University of Glasgow). Research interests include digital forensics, cybersecurity, residual data, and incident response. Has conducted research with Fortune 500 Organizations (financial services and manufacturing sectors), and law enforcement agencies.

Wayne Henry, USAF Lieutenant Colonel Henry is an Assistant Professor of Electrical Engineering at the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, Ohio USA. He received a B.S. in Computer Engineering from the Pennsylvania State University in 2004, and an M.S. in Computer Engineering in 2011 and a PhD in 2020 from AFIT. His research interests include cyber security, malware analysis, human-machine interaction, and information visualizations.

Daniel Farber Huang is a cybersecurity author, Consulting CISO, and advocate for online privacy. He has worked closely with numerous federal, state, and local law enforcement agencies across the U.S. He is a strategic consultant and advisor to a broad array of companies and organizations, ranging from entrepreneurial start-ups to multi-national corporations.

Gazmend Huskaj is a Doctoral Student in Cyberspace Operations at the Swedish Defence University. He received his MSc in Information Security from Stockholm University in 2015 as a distinguished graduate. Previously, he was Director Intelligence in the Swedish Armed Forces focusing on cyber-related issues. He is also a ISACA Certified Information Security Manager (CISM).

Richard Adeyemi Ikuesam an active researcher pioneering a digital policing and forensic project for developing nations, using Nigeria and South Africa as a hub for West Africa and Southern Africa respectively. Received his M.Sc. and Ph.D. in Computer Science with distinctions from the Universiti Teknologi Malaysia. Currently Assistant Professor (Cyber Security section of the IT department, Community College, Qatar).

Bilge Karabacak worked at a research institute for 14 years. He earned his M.S. in Computer Science and Ph.D. in Information Systems. He co-founded a cybersecurity startup in California. After three years in Silicon Valley, he started working for a multinational bank as a security analyst. Currently, he is a full-time faculty at Franklin University.

Saltuk Karahan is a Lecturer in Old Dominion University's Department of Political Science and Geography. He is also the Program Coordinator for ODU's School of Cybersecurity. Dr. Karahan's research interests and projects focus on the central themes of interdisciplinary approach to cybersecurity education and how the developments in information security technology impact international security.

Mutomb Japhet Kayomb is a junior lecturer of information technology at Cape Peninsula University of Technology, SA. He is completing his Master's study in Information and communication technology in the stream of communication networks. His main research area is computer security.

John Kiger currently is a graduate Cybersecurity student at Rowan University. In 2020 he received his B.S. in CS from Rowan University (Honors). In 2021, he was the recipient of the National Science Foundation's Scholarship for Service Award. His areas of interest focus on automated threat mitigation and malware analysis.

Neal Kushwaha is a recipient of the Silver Medal of Bravery and commendation from the Governor General of Canada. He advises and guides private and public organisations with their cyber challenges as it relates to cybercrime, espionage, sabotage, and warfare. He has written and published several papers on the intersection of cyberspace and International Law.

Martti Lehto, (Military Sciences), Col (GS) (ret.) works as a Professor (Cyber security) in the University of Jyväskylä. He has over 30 years' experience in C4ISR Systems in Finnish Defence Forces. He is Adjunct professor in NDU in Air and Cyber Warfare. He has over 190 publications. Since 2001 he has been the Editor-in-Chief of the Military Magazine.

Dr. Kristen K. Liggett is a Principal Human Factors Engineer in the Air Force Research Laboratory's 711th Human Performance Wing where she applies the science of human factors to Air Force challenges. Her current research focus is on the design and evaluation of user interfaces and cognitive work aids for a range of cyber operators.

Banks Lin, USSF. Captain Lin (BS, San Jose State University) previously served as cybersecurity test lead at the Air Force Operational Test & Evaluation Center, conducting operational-realistic cyber assessments on space and missile weapon systems. He now serves for the United States Space Force and is a student at Air Force Institute of Technology studying for a Master of Science in Cyber Operations.

Christoph Lipps, M.Sc., graduated in Electrical and Computer Engineering (University of Kaiserslautern, Germany, where he lectures. A Researcher and Ph.D. candidate at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, research focuses on Physical Layer Security (PhySec), Physically Unclonable Functions (PUFs), Artificial Intelligence (AI), entity authentication and all aspects of network and cyber security.

Ran Liu is a Ph.D. student in University of Maryland, Baltimore County, advised by Professor Charles Nicholas. He received his BS degree at Wuhan University, and MS Degree from the University of California, Santa Barbara. His primary research interest includes machine learning, quantum computation and malware analysis.

David Long is a sec Dev ops engineer working at Facebook. He holds an MSc. in Applied Cybersecurity from the Technological University Dublin, Ireland. His research interests lie in Malware classification and Deep Learning.

Clara Maathuis is Assistant Professor in AI and Cyber Security at Open University in the Netherlands. With a PhD in AI and Military Cyber Operations from Delft University of Technology, she is involved in teaching different AI and cyber security courses and conducts research in AI, cyber/information operations, military technologies, and social manipulation.

Bill Mahoney is the Charles and Margre Durham Distinguished Professor in Cybersecurity at the University of Nebraska at Omaha, where he teaches both undergraduate and graduate cybersecurity classes. His research areas include code obfuscation, reverse engineering and anti-reverse engineering techniques, and vulnerability analysis. An additional interest is in vulnerabilities with critical infrastructure equipment.

Arianit Maraj is a recognized cybersecurity expert and Professor at the University level. From 2005, he worked for Telecom of Kosovo. He is Doctor of ICT Engineering. In 2018, he performed Postdoctoral Research studies at Florida Institute of Technology-USA, as a Fulbright Scholar. His research interests lay in Data security, Networking and Wireless communications.

Michael Motlhabi received his BSc Hons in Computer Science (Cum Laude 2011) and his Master of Science degree from the University of the Western Cape. He worked in the Telecommunications industry as a Packet Optical Transport Network engineer for 6 years. He is currently a Senior Cybersecurity Specialist at the Council for Scientific and Industrial Research (CSIR).

Qazi Muhammad Ali. I am a PhD Scholar at Superior University Lahore, Pakistan. I am perusing my PhD in Business Administration. I had attended many conferences in my research life. My main research area is on Behaviours and Human Resource Management.

Tristen Mullins is a PhD candidate in Computing and a National Science Foundation CyberCorps Scholarship for Service student at the University of South Alabama. She has a BSc in computer engineering and her research interests include side-channel analysis, countermeasure design, and embedded systems security.

Petr Munster is Associate Professor of fiber-optics at the Brno University of Technology, Department of Telecommunications, Czech Republic. He received his Ph.D. degree in fiber-optic networks in 2014. His research interests include fiber-optic sensors, especially distributed fiber-optic sensors, security of fiber-optic infrastructures, and fiber-optic data and non-data transmission.

Calvin Muramoto is a graduate student pursuing a master's degree in cyberspace operations at the Air Force Institute of Technology, USA. Mr. Muramoto has a B.S. in Computational Modeling and Data Analytics with a focus in Economics from Virginia Tech. His research focus includes machine learning, cyber-physical systems, and cryptology.

Bhagawan Nath, is a PhD candidate in the University of Jyväskylä, and a lecturer in the faculty of Information Technology, Ton Duc Thang University. He has taught various IT and Computer Science related courses at different universities earlier and his research interests include Artificial Intelligent, Deep Learning, Network Security and Cyber Security.

Tabisa Ncubukezi is a communication networks lecturer at the Cape Peninsula University of Technology. She is a doctoral candidate at the same institution. Her primary research areas are communication networks, cloud computing, cyber security, and e-Learning.

Brian Nussbaum is an assistant professor in the College of Emergency Preparedness, Homeland Security and Cybersecurity (CEHC) at the University at Albany. His research has been published in the Business Horizons, the Journal of Cyber Policy, Computer Law & Security Review, and the Journal of Financial Crime, among other venues.

Jacob Oakley is a cybersecurity expert and author with over 15 years of experience improving strategic, enterprise-level security architectures for threat and risk management of globally distributed Department of Defense (DOD) and Fortune 500 networks. Additionally, he has served as the technical principal for multiple cybersecurity R&D efforts.

Ikuomor Mabel Ogiriki received her B.S in computer science from Lincoln University, Pennsylvania. She is currently pursuing her M.S degree in computer science at Rowan University, New Jersey. Being a teaching fellow with Rowan University and front-end software engineer for 3 years is also part of her work experience. Her area of interest includes pursuing a career in the field of cyber security, software security, and design.

Xavier-Lewis Palmer is a Biomedical Engineering PhD Candidate Old Dominion University and recent cybersecurity Masters graduate. His dissertation-based work is at the intersection of cancer biology and tissue engineering, but his entire body of work is interdisciplinary, aiding contributions to discussions and work on intersections that concern healthcare, security, and the community.

Johanna Parviainen, student of Security Management at Laurea University, works as a Detective Inspector at local police. She has years' experience of leading different crime investigations including crimes against children, last 2.5 years she has worked with economic crimes. Her Master Thesis explores sexual abuse of children online and CAM investigations in Finland.

Bianca Pasikhani holds a masters in information assurance from University of Maryland Global Campus (UMGC) and an undergraduate degree from Columbia University in Political Science. She is interested in artificial intelligence and cloud technology. The industries she is focused on includes construction especially the construction cloud. She is also interested in the overlap of information assurance and project management. She currently works in the construction industry for a global general contractor.

Matthew Peterson is a PhD. Candidate at the University of South Alabama. He currently works professionally as a Software Engineer in the medical industry. Insights gained from working in the industry gave him speaking opportunities at conferences such as the Gulf Coast Code Camp. His primary interests in the computing field are Blockchain, machine learning, and information security.

Heloise Pieterse is currently employed as a senior researcher within the Cyber Warfare research group at the Council of Scientific and Industrial Research. She completed her PhD Computer Science degree in 2019, with a focus on identifying the authenticity of smartphone data. Her interests include digital forensics, mobile device security and cyber security.

Jordan Plotnek received the B.Eng degree with first class honours in telecommunication engineering from Swinburne University of Technology, Melbourne, Australia in 2013. He is currently a Ph.D. Candidate at the University of South Australia and co-Founder of Anchoram Consulting. His research focuses on cyber-physical conflict, emerging threats, and resilience, particularly in relation to space systems.

Ruan Pretorius. I live in Cape Town, South-Africa, and I am 24 years old. I am currently completing my master's degree in Socio-Informatics at Stellenbosch University while being employed as a Systems Analyst. My tertiary education includes my bachelor's degree in Information Systems, and my honor's degree in Socio-Informatics, for which I achieved a cum laude.

Jyri Rajamäki is Principal Lecturer in Information Technology at Laurea University of Applied Sciences and Adjunct Professor of Critical Infrastructure Protection and Cyber Security at University of Jyväskylä, Finland. He holds D.Sc. degrees in electrical and communications engineering from Helsinki University of Technology, and PhD degree in mathematical information technology from University of Jyväskylä.

Caleb Richardson is master's student studying computer science at the Air Force Institute of Technology, USA. He received his bachelor's in computer science from the Air Force Academy in 2021. His main research area is cyber security.

Nicole Rothwell is an Honours student in Information Systems Management at Stellenbosch University, having graduated with a Bachelor of Arts undergraduate degree in Socio-Informatics in 2020.

Char Sample is the Chief Research Scientist for the Cybercore division at the Idaho National Laboratory, where she oversees and guides research into all areas of cybersecurity within the division. With over 25 years of experience her research areas are data resilience, AI/ML, ICS/CPS and wireless, fake news, cultural values in cybersecurity.

Keith Scott is Programme Leader for English Language at De Montfort University in Leicester. His research operates at the intersection of communication, culture and cyber, with particular interests in influence, information warfare, cybernetics, and simulations and serious gaming as a training, teaching, and research tool.

Syed Shahzad. PhD student (Australian Defence Force Academy, ADFA), University of New South Wales, Australia. Received his postgraduate degrees in computer science from ADFA and finance from St Hugh's College, University of Oxford. He is a cyber security practitioner in the Defence industry. Main research areas are cyber resilience, cyber security, and risk management for critical space infrastructure.

Michael Shepherd is a junior at Wentworth Institute of Technology working towards his bachelor's degree in computer science. During the summer of 2021, he worked as an intern under the Trusted and Resilient Systems department at Riverside Research, where he studied and developed transient execution and side channel exploits.

Jantje Silomon joined the Institute for Peace Research and Security Policy at the University of Hamburg as a researcher in 2019, moving to the newly founded "International Cybersecurity" team in 2021. Previously, Jantje conducted her doctoral research on the topic of malware weaponisation at the University of Oxford.

Jussi Simola is a doctoral researcher of cyber security at the University of Jyväskylä. His expertise includes decision-support technologies, Situational Awareness Systems, information security, and continuity management. His current research work is focused on the development of Maritime Cyber Security. He has been involved in the development of a common EU early warning system.

Avinash Singh is an emerging researcher (BSc Hons and MSc in Computer Science with distinctions from the University of Pretoria). Currently a lecturer pursuing a PhD (Department of Computer Science). He has published in international conferences and journals. He is also head of the Intelligent Cyber Forensic Lab (University of Pretoria) and a member of the DigiForS research group.

Siphesihle Sithungu is from Johannesburg, South Africa and holds a MSc in Computer Science (University of Johannesburg). He is a lecturer at the University of Johannesburg and his research interests are artificial intelligence and critical information infrastructure protection. Mr. Sithungu is a technical committee member for the International Conference on Computational Intelligence and Intelligent Systems.

Meghan Strang is a Human Factors Engineer in the Air Force Research Laboratory's 711th Human Performance Wing. She received her Masters of Science in Engineering Management from the Air Force Institute of Technology in 2020. Her main research areas are human system integration, decision making, and the evaluation of operator interfaces.

Sneha Sudhakaran is a faculty candidate from Louisiana State University, LA. She is expected to receive PhD in Computer Science from Louisiana State University in 2022. She is an active member in contributing her research work to multiple conferences and journals. Her research interest includes Memory Forensics, Android Application Security and Reverse Engineering.

Jeffrey C. Sun is Professor of Higher Education & Law, Distinguished University Scholar, and Associate Dean for Innovation and Strategic Partnerships at the University of Louisville. In addition, Dr. Sun has served as Project Director and Principal Investigator for over \$19 million in grants and contracts for projects advancing professional/career education and higher education policy and law.

Dr. Jim Sweeney is a cybersecurity engineer working with Exida Technologies LLC and is a part time paramedic. He has over 25 years of experience in emergency management, risk assessment and analysis. He has worked in both the public and private sector and has held multiple roles during incident response.

Dr. Vu N. Tran is a computer scientist in the Navy and a part-time faculty member of Capella University. Dr. Tran has over 35 years of working experience, including 20 years in security engineering in the satellite and cable broadcasting industries in North America, Europe, the Middle East, and Asia. Dr. Vu N. Tran is a computer scientist in the Navy and a part-time faculty member of Capella University. Dr. Tran has over 35 years of working experience in both the private and government sectors.

Stephen Taylor is a Professor of Computer Engineering at Dartmouth College, and a co-owner in Web Sensing LLC, an R&D company specializing in network security solutions. His academic research focusses on systems security using System-on-Chip and FPGA devices. He is a former DARPA Program Manager and member of the US Air Force Scientific Advisory Board.

John Theborge is completing his Masters degree in Cyber Operations at the Air Force Institute of Technology, Wright-Patterson AFB, OH, USA. Research interests include cyber security, space systems, and offensive cyber operations.

Christina Thorpe, BSc, PhD Christina Thorpe graduated with a B.Sc. (Hons) in Computer Science from University College Dublin in 2005 and a Ph.D. in Computer Science in 2011. She was a postdoctoral research fellow in the Performance Engineering Lab in UCD from 2011 - 2018. She is currently a Lecturer in Cyber Security in the Technological University Dublin.

Maija Turunen is a Military Science student at the Finnish National Defense University. Her main research areas consist of cyber warfare, cyber deterrence and Russia.

Gilliam Van Der Merwe is a postgraduate student of Information Systems Management at Stellenbosch University, SA. He is passionate about the realm of cybersecurity. He has a keen interest in improving his understanding and insight into cyberthreats and the techniques they use. His goal is to improve prediction and mitigation in cloud infrastructure environments.

Stephanus Petrus van der Walt is currently a solutions architect in cybersecurity with over 20 years of experience with numerous industry certifications such as 3 x Cisco CCIEs. He holds a BSc degree in Computer Science and an Honours BSc in Computer Science degree from University of Stellenbosch. He is currently an MSc student at the University of Pretoria.

Namosha Veerasamy has a BSc:IT Computer Science, a BSc: Computer Science (Honours Degree) and MSc:Computer Science (distinction) (University of Pretoria). Holds a PhD (University of Johannesburg). Currently a senior researcher at the Council for Scientific and Industrial Research in Pretoria. A Certified Information System Security Professional and Certified Information Security Manager. Involved in cyber security research/ governance for over 10 years.

Gábor Visky is a Researcher at NATO Cooperative Cyber Defence Center of Excellence, focusing on Industrial Control Systems. His prior assignments include 15 years of designing hardware and software of embedded control systems and researching its vulnerabilities. As a PhD student at Tallinn Technical University, Gabor researches the cyber security of modern ships

Prof. Suné von Solms is an Associate Professor and Head of Department at the Faculty of Engineering and the Built Environment at the University of Johannesburg, South Africa. She obtained a Ph.D in Computer Engineering, is a registered professional engineer with the Engineering Council of South Africa (ECSA) and a National Research Foundation (NRF) rated researcher. Her research interests

include networks and communication, engineering education, the social and human aspects of engineering and cybersecurity. She is actively involved in engineering and community engagement projects within rural communities. Suné is also involved in research relating to cybersecurity-related skills and competency development of engineers.

Murdoch Watney is a professor at the University of Johannesburg, South Africa. She teaches criminal law. Her field of interest is cybercrime and cybersecurity. She has published in textbooks on criminal law and has delivered peer-reviewed papers at national and international conferences.

Cyberbullying Indicator as a Precursor to a Cyber Construct Development

Salam Khalifa Al-Romaihi¹ and Richard Adeyemi Ikuesan²

¹Community College Qatar, Doha, Qatar

²Zayed University, Abu Dhabi, UAE

Salman63962@ccq.edu.qa

Richard.Ikuesan@zu.ac.ae

Abstract: The current global pandemic occasioned by the SARS-CoV-2 virus has been attributed, partially, to the growing range of cyber vises within the cyber ecosystem. One area of such impact is the increasing tendencies of cyber-bullying among students. Cyberbullying -the act of subjugating others using a cyber platform- is a growing concern among educators, especially in High-Schools. Whilst studies have been carried out towards understanding this menace, the approach towards identifying indicators of cyberbullying is largely missing in the literature. To address this research gap, this study proposed a cyberbullying framework based on the identification of some observable behavioral indicators. Using a self-administered measurement instrument from 30-respondents, the study observed the probability of a cyberbully construct, as a potential measure of the presence of cyberbullying; a probability that has been largely ignored in extant literature. This observation presents a veritable tool for the development of an active and integrated learning platform void of abuse among students. Furthermore, within the cyber education ecosystem, a cyberbullying construct would provide a mechanism for the development of an appropriate online learning platform, which would be useful to the information system and cyber education research communities.

Keywords: Cyberbullying indicator construct, Cyberbullying, cybercrime, Olwens bullying cycle, behavioral indicators

1. Introduction

Like physical bullying, cyberbullying can occur in diverse ways, and can also reflect in myriad forms, ranging from direct to indirect, intentional to unintentional, one-to-one, one-to-many, and many-to-many as well as overt-to-covert actions. However, unlike physical bullying, cyberbullying can transcend borders, boundaries, and even temporal-spatiality with a long-lasting effect. For instance, any act posted online could be retained for years until such an act is repealed or taken down. The concept of cyberbullying, as highlighted in Corliss, (2017) and Lee and Chun, (2020), is rooted in diverse ontological definitions which primarily involve the use of technology. Put succinctly, cyberbullying is the act of leveraging the power of technology to cause harm or discomfort to another. This act is often associated with frequent or repeated communication occasioned by aggressive tendencies (towards a hurtful, disgraceful, harmful, or discomfort outcome) to recipients (Nakano *et al.*, 2016; Buelga *et al.*, 2020). Redmond, Lock and Smart, (2020) further dissected technology used to include electronic mails, telecommunication text messages, chat rooms, mobile phones, web, and mobile cameras, as well as websites. With the steady growth and adoption of technology, the cyberbullying act can be projected to increase. Different dimensions of cyberbullying acts have been identified in the literature (Langos, 2012). Cyber personality has also been defined in existing studies (Adeyemi *et al.*, 2016; Adeyemi, Abd Razak, *et al.*, 2017; Adeyemi, Razak and Salleh, 2017; Adeyemi, Razak, *et al.*, 2017; Ikuesan *et al.*, 2019) with contradictory cyber:physical personalities (Gardini, Cloninger and Venneri, 2009; Mohammadi and Vinciarelli, 2015; Ikuesan, Abd Razak and Salleh, 2016). The next section explores the different dimensions of cyberbullying with a view to highlighting the underlying commonalities.

2. Cyberbullying Dimensions

In trying to understand the different dimensions of cyberbullying, it is essential to fully understand what cyberbullying is. Within the research community, cyberbullying has garnered diverse interpretations and consequently, different meanings (Langos, 2012; Redmond, Lock and Smart, 2020). Building on traditional bullying, as measured by the Olwens Bullying questionnaire, studies have revealed that respondents often equate cyberbullying to acts that involve the internet and technology (Dennehy *et al.*, 2020; Fernández-Antelo, Cuadrado-Gordillo and Parrà, 2020). For example, the item "somebody used the internet or a cellphone to harm/offend me" has been regarded as an act of cyberbullying. As elaborated in existing studies (Corliss, 2017; Redmond, Lock and Smart, 2020), these definitions and perspectives, as a construct, present a complex phenomenon that requires more insight. Position in Langos, (2012) summarized the existing definitions within the context of intention: direct or indirect. The study coined the definition as "*Cyberbullying involves the use of ICTs to carry out a series of acts as in the case of direct cyberbullying, or an act as in the case of indirect*

cyberbullying, intended to harm another (the victim) who cannot easily defend him or herself". The definition reveals cyberbullying to include acts of willfulness, power imbalance, repetitiveness, perception-based cognition, as well as a technology-integrated process. Moreover, the integration of "cyber" into traditional bullying presents a susceptibility and vulnerability beyond temporal-spatial boundaries, which further enhanced the proclivity for bullying. Therefore, irrespective of the perspective of the definition from diverse stakeholders, the underlying tenet of cyberbullying remains visible for evaluation. Evaluating cyberbullying, its effect, and potential mitigation approaches can be hinged the process of identifying cyberbullying cues, hereinafter refers to indicators. A study in (Redmond, Lock and Smart, 2020) presented three perspectives of cyberbullying indicators based on bully attributes, using a taxonomical structure. The study leveraged bully attributes, bullying categories, awareness based on the victim's and educator's perspective. These attributes, further classed as Category-1, provide a baseline for cyberbullying identification within the cyberbullying conceptual framework. Whilst the identification of cyberbullying act, specifically from the student's perspective, presents a means for cyberbullying construct development (Dennehy *et al.*, 2020; Fernández-Antelo, Cuadrado-Gordillo and Parrà, 2020), the content presented in the conceptual framework lacks useful composition that can be used as indicators or precursors. Cyber bully indicator can be defined within a context or in perspective: the victim, perpetrator, or collaborators, for example. These perspectives are further distilled using the modified Olwen's bullying cycle shown in Figure 1. Two potentially applicable perspectives (E2 and E3) are integrated into the cycle: unwilling supporter and unconscious supporter. These perspectives has been clustered together in some studies to mean unintentional bullies (Langos, 2012). However, recent advances in technology have provided a baseline for the separating these two perspectives.

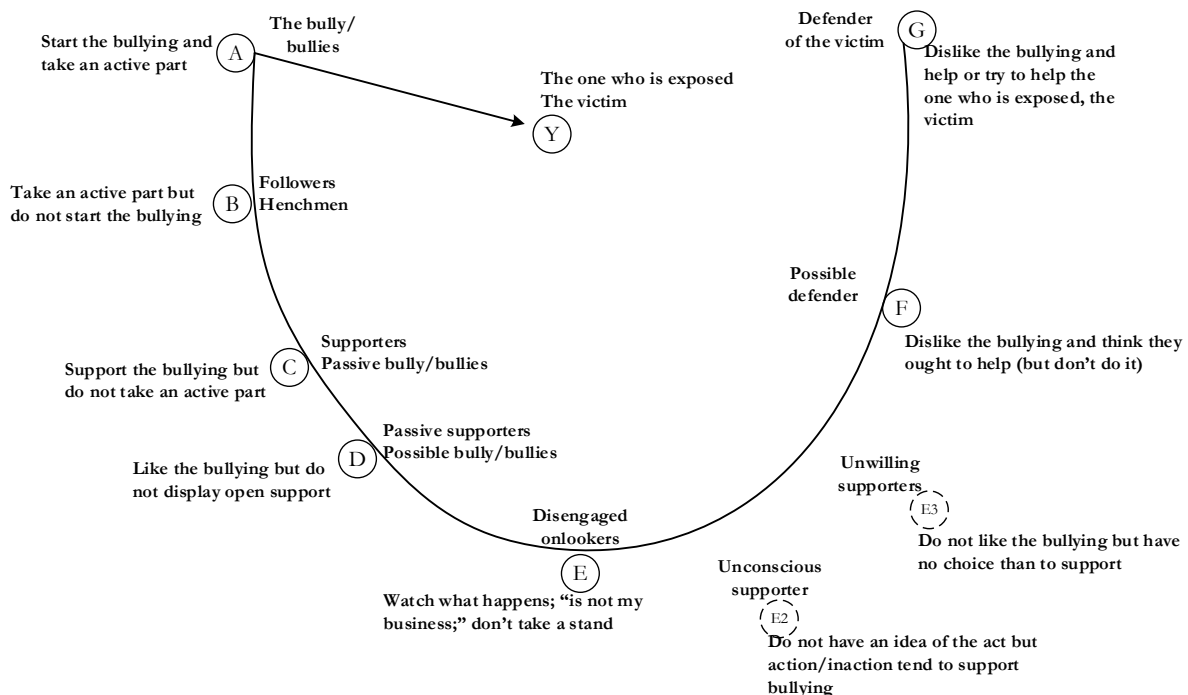


Figure 1: Modified Olwen's bullying cycle.

An unwilling supporter is an entity who does not like bullying but thinks he/she has no option other than to partake. This entity can also be termed a victim of the bullying act. The second category; an unconscious supporter, is an entity whose action/inaction unknowing encourages bullying. This category could potentially form a large population. This is particularly applicable to a cyber-induced bullying process where entities unknowingly provide support to abusive acts that could harm others. Therefore, approaches that provide measures for reliably identifying signs and markers of cyberbullying are largely missing in the literature. Indicators in this regard refer to a measurable instrument that can be used to deduce the tendency and or presence of cyberbullying act/process. To address this observed limitation, this study explores the probability of extracting inherent markers of cyberbullying. The study further proposed an identification mechanism rooted in the result of the exploratory process. To the best of the Authors' knowledge, this study serves as a pioneering insight towards the development of a reliable construct, hereinafter referred to as the Cyber Bullying Indicator construct (CBIC). Within the general body of information systems, a construct is a concept of abstraction that cannot be directly described using a singular entity, however, exist, nonetheless. Constructs such as Perceived

Ease of Use (PEU), Perceived Usefulness (PU), and Perceived Behavioral Control (PBC) are examples of constructs that have been widely used in the literature. This study, therefore, presents a foundational basis on which a CBIC can be developed. This is achieved based on an underlying research question given thus: *Given the growing tendency of online cyberbullying in educational settings, can the use of a quantitative exploratory study be used to identify any underlying constructs that can explain cyberbullying with a Cronbach's alpha reliability greater than 0.7?*

The remainder of the manuscript is organized as follows: the next section discusses the methodology applied in this study to explore the conceptualization of the construct. This is followed by the presentation of the exploratory result and an in-depth analysis of the result. Conclusion and future works are thereafter provided.

3. Methodology

Given the nature of cyberbullying, and the scope of this study, an exploratory quantitative approach was considered. A set of measurement instruments was developed based on the exploration of common cyberbullying concepts used in literature. The final compilation comprised 20-items. Each item represents a potential measure of bullying indicator. This includes the two ends of the bullying cycle spectrum identified in Figure 1: the bully (A) and the victim (Y). To select participants for this study, three criteria were defined as shown in Table 1. Taken together, the criteria provide a guide for respondent selection.

Table 1: Respondent Inclusion Criteria

Criteria	Description
1	Respondents must be in academics in departments that work with information technology either for duties or develop information technology content.
2	Respondents should have knowledge of cyberbullying, either from a behavioral science or experience. This can be within the social science discipline.
3	Respondents must be a parent who has a child(ren) in secondary school(s)

Upon receipt of ethical clearance from the institution, the preliminary exploratory process was conducted. The questionnaire instrument was sent to selected Faculties in the institution via email. This process leveraged the purposive non-probability sampling technique as opposed to the probabilistic random sampling technique. The summary of the measurement instrument developed for this study is further shown in Table 2. The instrument contains 20 items adapted from existing studies, which can be further grouped into two classes.

Table 2: Measurement Instrument for CBIC

Item	Content	Factor
Q3	Abruptly shutting down or walking away from the computer mid-use	Phase-1
Q6	Expresses abrupt reaction when his/her online-related gadget is being touched	
Q9	Upset or frustrated after going online or gaming	
Q12	Withdrawing from friends and family in real life after a long hour of techno-isolation	
Q14	He or she clears the computing device screen abruptly when someone enters the room.	
Q16	He/she starts becoming secretive about their Internet activity.	
Q18	Does not like to use websites they used to spend time on.	
Q19	He/she seems upset, highly irritable, or emotional after being on the computer	
Q5	Unexplained anger or depression, especially after going online	Phase-2
Q8	Sudden detachment from electronic appliance	
Q11	Demonstrates uneasy, nervous, or scared when he/she receives an electronic message	
Q17	There is a sudden drop in computer use and seems to avoid its use	
Q20	He/she seems upset, highly irritable, or emotional after reading their text messages or email on their mobile device	

A total of 30 responses were received. Whilst the sample size of 30-response is relatively small, the authors assert that the sample presents a baseline for further studies. As highlighted in Lee and Chun (2020), studies on conceptualizing cyberbullying can leverage a sample size greater than 10-response. The sample in the current study is relatively higher than the sample used in Phase-2 of the concept mapping process in Lee and Chun (2020). Given that this is the preliminary insight into the subject matter, the sample size was considered sufficient for this study. Statistical Package for Social Science (SPSS) v23 was used to conduct the exploratory process while SMART PLS (student License) was used to explore confirmatory factor analysis. The detail of the result is presented in the next section.

4. Result and Analysis

An exploratory analysis typically involves the process of extracting the underlying intrinsic cluster in data. To do this, the Cronbach's Alpha test of instrument reliability was carried out. The result showed that the items were indeed reliable with a Cronbach's Alpha value of 0.803. The Kaiser, Meyer, Olkin (KMO) measure of sampling adequacy generated poor sphericity at 0.346 which was an expected outcome, given the overall sample size used in this study. In the initial factor exploration, no pre-defined number of clusters was considered (which implied an Eigenvalue > 1 was considered). A total of 5-underlying factors was extracted using maximum likelihood as the method of extraction. A further observation of the result showed that the data has two underlying factors. Thus, pruning the factor into two and suppressing coefficients < 2 , the pattern analysis revealed two overlapping factors, at a $\chi^2/df = 1.612$. From the extracted pattern, 4-items had relatively similar cross-loading. These items were deleted. Furthermore, 7-items and 9-items were classified as factors 1 and 2 respectively. To perform the confirmatory analysis, the pruned data was used to model the path analysis. The path model revealed the factor loading of three items to be less than 0.4. These items were further removed. The resultant path model is presented in Table 3 and Figure 2. The observed path coefficient between the two factors was 0.55.

Table 3: Construct Reliability and Validity

Factor Description	Cronbach' Alpha	Rho Alpha	Composite Reliability (CR)	Average Variance Extracted (AVE)	R Square
Phase-1	0.831	0.842	0.87	0.463	
Phase-2	0.712	0.749	0.739	0.446	0.308

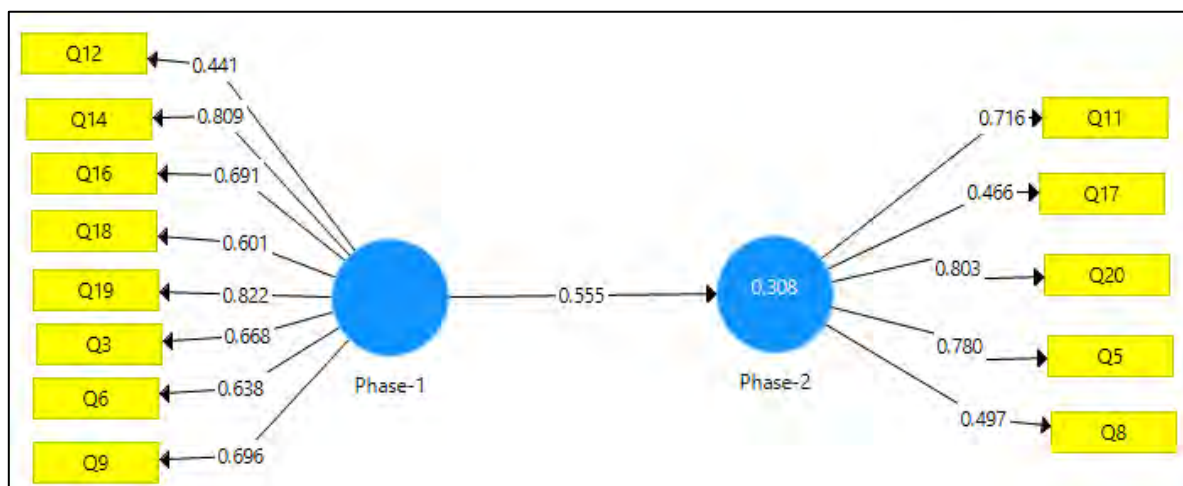


Figure 2: Path Model towards a confirmatory factor analysis

From the result shown in Table 3, it certainly appears that the two factors can model the measurement items. The factor loading for all items in Factor-1 (hereinafter referred to as Phase-1) and Factor-2 (hereinafter referred to as Phase-2) are greater than 0.4. In terms of construct reliability and validity, both factors satisfied the standard thumb rule for the Cronbach's alpha, rho alpha, and composite reliability. However, the AVE was below the 0.5 thumb rule. Using Phase-1 as a predictor variable for Phase-2 and vice versa, the study generated an identical regression weight (0.555) which further shows the probable existence of two variables that can be used to measure the proposed cyberbullying construct.

5. Discussion

The result presented in the previous section revealed the probability of the existence of a cyberbullying indicator construct. A cyberbullying construct provides a baseline for measuring the presence of bullying, which can be measured by parents, guardians, as well as educators. Often, subtle bullying process among secondary school students can go undetected until it causes damages to the victims. This study, therefore, serves as a first step towards mitigating cyberbullying act, especially among secondary school students. With such an indicator, different stakeholders can identify signs before it degenerates into psychological and physical trauma. This is especially useful given the trendy adoption of digital lifestyle occasioned by the Covid-19 pandemic and the potential for an inseparable work-life balance for parents. Cyberbullying is relatively difficult to identify without

the core indicators, relative to physical bullying (Lee and Chun, 2020). Indicators in this regard reveal necessary precursors that can be used to glean the emergence or inception of a cyber induced bullying process. A summary of the final measurement is given in Table 2. A breakdown of the item's clusters in Table 2 shows that there exist some underlying tendencies that can be further classified into Phase-1 and Phase-2 behavior. Both Phases 1 and 2 depict a reflective indicator that shows the behavior of an entity after an incident has occurred. For example, the concept of abrupt change in behavior after an online experience could reflect what had transpired during an Internet surfing period. This behavior could result in excitement on the part of the bully, while the converse would be seen on the victim. The items defined within this study can be further examined in line with the underlying tenet in each item. Though words like withdrawer, abrupt, highly irritable, emotional, as well as upset, are general terms associated with victims of a bully, the concept of cyber content is revealed in words like an electronic message, internet gaming, online activities, techno-isolation, computing device, and online-gadget. These keywords can be associated with cyber activities. Though the current exploratory analysis generated two clusters, a confirmation of this would be rigorously carried in the subsequent study. This is essentially required to formulate a steady construct for the CBIC. A theoretical analysis of the measurement items in each Phase did not reveal a remarkable distinction. This further explains the need for further study.

Ideally, a construct consists of a singular class, and the process to formally develop such would pass through a series of reliability tests. Further study with a significantly larger sample size will be carried out to ascertain the feasibility of developing the CBIC. As an ongoing work, a further study that involves the development of a nomological network, as well as external factors that can be used to validate the probability of the existence of the CBIC, will be carried out. This attempt will leverage constructs from other information systems models such as the theory of planned behavior and the extended technology acceptance model. Also, other cyberbullying constructs such as the CYB-AGS (Buelga *et al.*, 2020) aggressive scale can be explored. Such a network can be used to measure the impact of cyber education and awareness against potential bullies. One aspect that is not considered in this study is the influence of the internet as a tool that aids anonymity. This notion can have a potential impact on who becomes a bully, and the transformative tendencies of cyberbullying. For instance, some studies (Haddadain, Abedin and Monirpoor, 2010; Yue *et al.*, 2010; Sari, 2016) have shown that the anonymity provided by the Internet induces different kinds of personalities on the Internet. This could further influence the degree of power imbalance associated with bullying. Opinion in Dennehy *et al.*, (2020) argued that the bully's "inability to witness the victim's reaction may diminish their empathetic response potentially leading to more harshness in the cyber behaviors". In addition, online anonymity could induce unintentional bullying, which existing studies have not explored. Online anonymity can be perceived differently between the bully and the victim. Further study on the reaction of Victims to cyberbullying is an aspect that can provide concrete insight into this perspective. Conjecturing or attempting to hypothesize on the possible behaviorism of the victim without empirical proof, could be misleading.

6. Conclusion

The concept of cyberbullying has been transformed with the advances in technology. This study explored the potential of developing a construct that can be used by different stakeholders to pinpoint and or identify the inception of cyberbullying act. Using an exploratory process, the study observed with high reliability, the probability of the existence of such a construct, which is further defined as cyberbullying indication construct (CBIC). However, the sample size used in this current study presents a major limitation of the study. A CIBC, when fully explored on larger sample size, can be a veritable tool for mitigating cyberbullying among vulnerable entities. Such a tool can be leveraged by the wider scientific community in advancing cyberbullying research. Furthermore, the integration of the proposed construct with other information systems constructs can be used to develop a nomological network for studying the effect of cyberbullying on students, as well as workplace culture.

References

- Adeyemi, I. R. *et al.* (2016) 'Observing consistency in online communication patterns for user re-identification', *PLoS ONE*, 11(12), pp. 1–27. doi: 10.1371/journal.pone.0166930.
- Adeyemi, I. R., Razak, S. A., *et al.* (2017) 'High-level online user attribution model based on human Polychronic-Monochronic tendency', in *2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017*. Seoul, Korea: IEEE Comput. Soc, pp. 445–450. doi: 10.1109/BIGCOMP.2017.7881753.
- Adeyemi, I. R., Abd Razak, S., *et al.* (2017) 'Leveraging human thinking style for user attribution in digital forensic process'. Indonesian Society for Knowledge and Human Development.

- Adeyemi, I. R., Razak, S. A. and Salleh, M. (2017) 'Individual difference for HCI systems: Examining the probability of thinking style signature in online interaction', in *Proceedings - 2016 4th International Conference on User Science and Engineering, i-USER 2016*. Melaka: IEEE, pp. 51–56. doi: 10.1109/IUSER.2016.7857933.
- Buelga, S. et al. (2020) 'Cyberbullying among adolescents: Psychometric properties of the CYB-AGS cyber-aggressor scale', *International Journal of Environmental Research and Public Health*, 17(9). doi: 10.3390/ijerph17093090.
- Corliss, C. L. (2017) 'The established and the outsiders: cyberbullying as an exclusionary process', *PQDT - UK & Ireland*. Available at: https://search.proquest.com/docview/2001151872?accountid=10673%0Ahttps://ed-primo.hosted.exlibrisgroup.com/openurl/44UOE/44UOE_VU2??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+&theses&sid=ProQ:ProQuest+Dissertati.
- Dennehy, R. et al. (2020) 'Young people's conceptualizations of the nature of cyberbullying: A systematic review and synthesis of qualitative research', *Aggression and Violent Behavior*. Elsevier, 51(January 2019), p. 101379. doi: 10.1016/j.avb.2020.101379.
- Fernández-Antelo, I., Cuadrado-Gordillo, I. and Parrà, G. M. M. (2020) 'Determining Factors in the Perception of Cyberbullying in Victimized Adolescents: Psychoeducational Implications', *International Journal of Emerging Technologies in Learning*, 15(24), pp. 38–51. doi: 10.3991/ijet.v15i24.19309.
- Gardini, S., Cloninger, C. R. and Venneri, A. (2009) 'Individual differences in personality traits reflect structural variance in specific brain regions.', *Brain research bulletin*, 79(5), pp. 265–70. doi: 10.1016/j.brainresbull.2009.03.005.
- Haddadain, F., Abedin, a. and Monirpoor, N. (2010) 'Appraisal of personality, family structure and gender in predicting problematic use of Internet', *Procedia - Social and Behavioral Sciences*, 5(2), pp. 850–854. doi: 10.1016/j.sbspro.2010.07.197.
- Ikuesan, A. R. et al. (2019) 'Polychronicity tendency-based online behavioral signature', *International Journal of Machine Learning and Cybernetics*, 10(8), pp. 2103–2118. doi: 10.1007/s13042-017-0748-7.
- Ikuesan, R. A., Abd Razak, S. and Salleh, M. (2016) 'Understanding Online Behavior: Exploring the Probability of Online Personality Trait Using Supervised Machine-Learning Approach', *Frontiers in ICT*. Frontiers, 3, p. 8.
- Langos, C. (2012) 'Cyberbullying: The challenge to define', *Cyberpsychology, Behavior, and Social Networking*, 15(6), pp. 285–289. doi: 10.1089/cyber.2011.0588.
- Lee, S. and Chun, J. S. (2020) 'Conceptualizing the impacts of cyberbullying victimization among Korean male adolescents', *Children and Youth Services Review*. Elsevier, 117(July), p. 105275. doi: 10.1016/j.childyouth.2020.105275.
- Mohammadi, G. and Vinciarelli, A. (2015) 'Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract', *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 3(3), pp. 484–490. doi: 10.1109/ACII.2015.7344614.
- Nakano, T. et al. (2016) 'Analysis of Cyber Aggression and Cyber-Bullying in Social Networking', *Proceedings - 2016 IEEE 10th International Conference on Semantic Computing, ICSC 2016*, (X), pp. 337–341. doi: 10.1109/ICSC.2016.111.
- Redmond, P., Lock, J. V and Smart, V. (2020) 'Developing a cyberbullying conceptual framework for educators', *Technology in Society*. Elsevier Ltd, 60(October), p. 101223. doi: 10.1016/j.techsoc.2019.101223.
- Sari, S. V. (2016) 'Computers in Human Behavior Was it just joke ? Cyberbullying perpetrations and their styles of humor', 54, pp. 2008–2012.
- Yue, C. et al. (2010) 'Psychological influences of blogging: Blog use, personality trait and self-concept', *2010 IEEE 2nd Symposium on Web Society*. Ieee, pp. 71–76. doi: 10.1109/SWS.2010.5607475.

Education and Training Against Threat of Phishing Emails

Ladislav Burita, Ivo Klaban and Tomas Racil

Department of Informatics and Cyber Operations, University of Defence, Brno, Czech Republic

ladislav.burita@unob.cz

ivo.klaban@unob.cz

tomas.racil@unob.cz

Abstract: The research results published in this article are oriented toward two areas: phishing email analysis and education for defense against the threats of phishing emails. The first topic builds on previous research primarily by analyzing changes in captured phishing emails over an interval of 4 weeks, half a year after the previous experiment. In this section, a statistical survey of phishing emails from both experiments is carried out and emails are segmented into categories focused on business, charity, asset transfer, and fund offers. The results of both experiments are then compared and validated. Based on this comparison and validation, a conclusion is made on trends and development in the phishing email domain in the last half a year. The second focus of our research is analysis of the existing education and testing systems for phishing emails. Based on the results of the analysis, a suitable system for university education and training against phishing and other malicious email threats will be designed. There is also an analysis of existing systems for improving and testing users' ability to recognize and react to phishing emails. Based on our findings about these systems, our own system is proposed. An experiment is prepared on "self-service" testing of phishing email detection skills performed by students with their colleagues. Some activists were employed to assist with this experiment; they will operate and prepare the environment according to the processed scenario. All experiments must be completely safe and effective at the same time. The experiments will be evaluated and the experience used to develop the education and training system at the university.

Keywords: Phishing threats and attacks, analysis of phishing emails, education and training

1. Introduction

Cyber security is a relatively new discipline that is undergoing dynamic development. Until ten years ago, the systematic management of information security was the prerogative of selected areas of economic and public sector organizations that work with classified documents. The growing risk associated with cyber threats has gradually but irreversibly changed the behavior of all companies and organizations in relation to cyber security. Phishing is the most common type of social engineering. In phishing, an attacker poses as a trustworthy source in an attempt to have the victim release personal or private information. Spear phishing is a popular type of phishing attack where the attacker provides information to a few select targets rather than generic information for mass targets (Summer and Yuan, 2019).

Phishing emails are attempts by scammers to trick someone into giving out their personal details. These details can include name, address, phone number, and other personal information like passwords, credit card numbers, and bank account numbers. Every system is only as durable as its weakest link, and, unfortunately, the human factor is often the most vulnerable in the cyber security of companies and organizations. People working in cyberspace are increasingly the target of attacks by malicious actors. One technique that is becoming increasingly important is social engineering and phishing attacks. The most effective measure against these techniques is user education and training.

People are vulnerable to phishing attacks because spoofed websites look very similar to legitimate ones. A study (Dhamija, Tygar, and Hearst, 2006) showed that people have trouble identifying phishing sites even in tests in which they have been alerted about the possibility of such attacks. The aim of this article is to describe the results of the analysis of phishing emails, to describe research experiments on education and training against phishing attacks, and to suggest a useful educational system for detecting such attacks. Phishing emails were collected in two stages with a six-month time interval not only to compare results of their analysis but also to see if their content evolves over time. For proper design of content and methods of education and training, it is necessary to understand phishing emails well.

In the part of this article on education and training, the procedures of an experiment with existing systems for improving and testing users' ability to recognize and react to phishing emails are first described. A suitable education and training system for the university department and faculty will be designed based on the results of this experiment.

Our research to date has confirmed some basic characteristics of phishing emails that include impersonal greetings, suspicious URLs, unusual content based on the stated sender and subject, requests for urgent action, and grammatical errors or misspellings. The offer of a large amount of money for business establishment and charity reasons, gifts from fund, or money transfer can be considered as unusual content, too.

This paper is organized as follows. The introduction provides an explanation of phishing and its threats for email users. The literature review is oriented to the areas of phishing attacks and their effectiveness, defense against phishing threats, and education and training to detect them. This is followed by phishing email analysis, research on testing and teaching systems to protect against phishing attacks, and design of our own education and training system.

The literature review makes it possible to find out relevant themes in papers indexed on Scopus that have dealt with phishing. Explanation of phishing as the most common type of social engineering, examples of phishing attacks and their effectiveness, methods of defense against phishing threats, reaction of users to phishing emails and their ability to recognize them, and education and training to prevent these attacks are common themes in the literature.

2. The literature review

The literature review made it possible to find out what themes are interesting in papers indexed on Scopus have dealt with in phishing. Explanation of the phishing as the most common type of social engineering, examples of phishing attacks and their effectiveness, using the methods of defense against phishing threats, reaction of the users at the phishing mail and ability to recognize them, education and training preventing of these attacks are often themes in the analyzed papers of the literature review.

The study (Butavius, Parsons, and all, 2015) was looking for answer of the influence the social engineering techniques of authority, scarcity, and social proof influence users' judgments on safety links in email with respect to different type of email (genuine, phishing, and spear-phishing). In the experiment participated 121 university students, were used 12 emails prepared in consultation with university teachers. During the experiment, for security purposes, was access to internet disabled, and the phishing link was modified. In all phishing and spear-phishing emails, the displayed text for a link was a description such as "Click here" or "Take the survey" rather than the actual link, and participants were advised, both verbally and in writing at the start of experiment. Participants correctly determined that legitimate links in genuine emails were safe to click 77% of the time. However, in spear-phishing emails, where the link was always unsafe, they incorrectly judged the link to be safe 71% of the time. Almost nearly half the sample (45%) did not judge any of the links in the spear-phishing emails as unsafe. For phishing emails, the percentage of responses that incorrectly judged the link to be safe dropped to 37% and 10% of the participants did not judge any of the links in the phishing emails as unsafe. Participants who were less impulsive in decision-making were more likely to judge the links in phishing emails as more dangerous.

Recently, targeted attacks are drastically increasing in both individuals and companies. For technical countermeasure against such a targeted attack, various methods such as email/web contents analysis etc., are developed and realized. However, attackers precisely exploit the most vulnerable part in order to achieve their goals. Therefore, spear phishing against human user is employed for such attacks in a large proportion. Moreover, in order to increase success probability of such attacks, attackers often adopt social engineering technique. In this paper, we present a current effort of our research group on combating targeted attacks employing spear phishing with using social engineering, through user education. The paper (Takata and Ogura, 2019) presents relationship between human psychological characteristics and vulnerability against social engineering. The result can be used for testing whether a user has vulnerability on some social engineering technique, and the testing result can be utilized for countermeasure or training. Next is mentioned development of a web-based self-learning material for countermeasure against social engineering, which employs interactive motion picture contents.

In the study (Liu, Zhou, and Zhang, 2020) were systematically investigate the effects of demographic factors, such as gender, age, education level, and technical experience, on phishing vulnerability in the workplace. The results of data analysis reveal the significant effects gender influence and technical experience on attacked by

phishing emails and show interesting interaction effects between education and technical experience. The findings of study can help improve the effectiveness of anti-phishing education and training.

In the experiment (Parsons, McCormac, and all, 2013), 117 participants were tested in 50 emails whether the knowledge that participants are undertaking a phishing study impacts on their decisions. One-half of users were informed that the study is assessing the ability to identify phishing emails. Results indicated that the informed participants were significantly better at correctly distinguish of phishing emails and it took them longer, because it increase in the ability to discriminate between phishing and real emails. Interestingly, participants who had formal training in security of information systems performed more poorly overall. The results have implications for the interpretation of previous phishing studies, the preparation of future studies and for training and education programs. Study suggests that when people are prepared on phishing risks, they adopt a more careful approach to emails.

The aim of the paper (Rastenis, Ramanauskaite, and all, 2019) was to evaluate how vulnerable to phishing attacks are personnel in a higher education institution. This group of users from the society, perspective is very educated, wise and resistant to fake information. In order to control the ethical issues of the experiment that was executed with employees of a university. Authors were able to assure the confidentiality of their personal data and all society of the university was informed on planned security experiments and the results of them. All personal information collected during the experiment was destroyed after it ended. The experiment consisted of four phases: preparation, execution, data analysis, and post-action. During the preparation, a fake web page and e-mail were created. As the experiment was executed, the email text states that the employee can get a present from the university and its partners but needs to provide some personal data in provided web page form on where the present should be sent. There were at least nine places and elements, which could identify the e-mail as phishing attack. The final part was training participants of the experiment, because it was planned to repeat the experiment to analyze the effect of executed training.

User security education and training is one of the most important aspects of an organizations security posture. Using security exercises to reinforce this aspect is frequently done by education and industry alike; however, these exercises usually enlist willing participants. In the research (Dodge, Carver, and Ferguson, 2007) was taken the concept of using an exercise and modified it in application to evaluate a user's propensity to respond to email phishing attacks in an unannounced test. The paper describes the considerations in establishing and the process used to create and implement an evaluation of one aspect of information assurance education. The evaluation takes the form of a exercise, where we send out a phishing styled email record the responses.

The study (Gordon, Wright, and all, 2019) was oriented to understand the impact of a phishing training program on phishing click rates for employees at a US healthcare institution. The experiment was prepared with 2 groups of population: offenders and no offenders. Offenders were defined as those that had clicked on at least 5 simulated phishing emails and no offenders were those that had not. It was calculated click rates for both groups, before and after a mandatory training program for offenders was implemented. A total of 5416 unique employees received all 20 campaigns during the intervention period; 772 clicked on at least 5 emails and were labeled offenders. Only 975 (17.9%) of our set clicked on 0 phishing emails over the course of the 20 campaigns; 3565 (65.3%) clicked on at least 2 emails. There was a decrease in click rates for each group over the 20 campaigns.

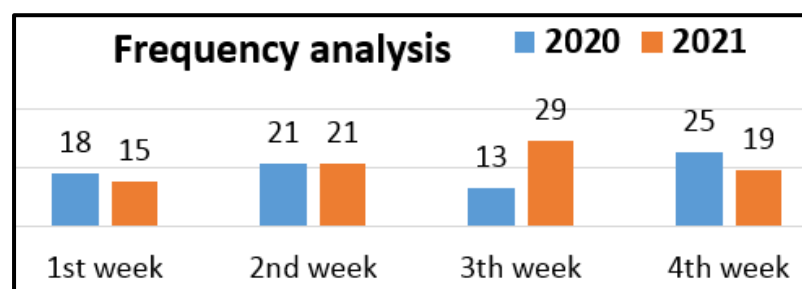
The paper (Kumaraguru, Sheng, and all, 2010) makes two research contributions. The first is a user study that demonstrates that existing anti-phishing educational materials are surprisingly effective if people actually read them. Educational and training materials were compiled from 24 various sources. The participants of experiment spent at most 15 minutes reading anti-phishing educational materials and then demonstrated significant improvements in their ability to recognize fraudulent websites when compared to a control group. The second contribution is an analysis of existing anti-phishing educational material. Users were given the following scenario: "You have received an email message that asks you to click on one of its links. Imagine that you have clicked on the link to see if it is a legitimate website or a spoofed website." It was then presented users with twenty websites and asked them to recognize whether a website legitimate is or phishing one. Twenty websites were divided into two groups with 5 phishing sites and 5 legitimate sites in each group. Analysis of training materials from company eBay, FTC, and Microsoft is in the Tab. 1.

Table 1: Result of training materials analysis (Kumaraguru, Sheng, and all, 2010)

Principle / Company	eBay	FTC	Microsoft
Multimedia	X	X	X
Contiguity	partially	---	partially
Personalization and story	---	X	---
Simplicity	---	X	partially

3. Phishing emails: the actual cyber threat

This section is focused on the analysis of phishing emails and their development. To effectively prepare university students and employees of companies and organizations to recognize phishing attacks, it is necessary that they become familiar with the content of phishing emails. To meet this goal, two experiments were performed in November 2020 and May 2021. In both, phishing emails sent to the working email box of one of the authors in a period of four weeks were detected, stored and then analyzed. An overview of the number of email messages is shown by the result of the frequency analysis in Fig. 1. The analysis shows that there was no significant change in the number of phishing emails sent.



General parameters of the phishing emails are given in Tab. 2. The parameter “Personal email” means that a person sent the email, while the parameter “Corporate email” means that a company sent the email. “Basic personal information” includes Name, Address, and Phone. “Detailed personal information” includes, in addition to basic personal information, such details as Occupation, Nationality, Date of Birth, etc. “Personal identification” is a requirement to disclose personal document number, passport number, letter of recommendation, etc.

Table 2: General parameters of the phishing emails

Message (77-2020/11; 84-2021/05)		2020/11		2021/05		TREND
		Sum	%	Sum	%	
1	Personal mail	54	70	51	61	<<
2	Corporate mail	23	30	33	39	>>
3	Contains the sender's name	73	95	77	92	<
4	Required basic pers-info	6	8	4	5	<
5	Required detailed pers-info	13	17	9	11	<<
6	Required pers-info fill in form	3	4	0	0	<
7	Required pers-identification	5	6	2	2	<
8	Required account number	3	4	4	5	=
9	Required a quick response	71	92	32	38	<<
10	Other language than english	5	6	10	12	>>
11	Money in mil. Usd	826		510		<<

Table 3: Categorization of the email message content

Message content	2020/11		2021/05		TREND
	Sum	%	Sum	%	
BUSINESS	21	27	25	29	=
Project, contract or investment	15	19	12	14	
Trading offer	2	3	10	12	
Job offer	4	5	2	2	
Payment for services	0	0	1	1	
CHARITY	17	22	6	8	<<
Charity due to a fatal illness	17	22	6	8	
FUND	14	18	22	26	>>
Gift, prize	4	5	2	2	
Fund of bank, Compensation fund	3	4	11	13	
Contractor of the fund; to share fund	4	5	7	9	
Inheritans	3	4	2	2	
TRANSFER	21	28	15	18	<<
Money from bank account to receiver account	19	25	15	18	
The pick up shipment	2	3	0	0	
OTHER	4	5	16	19	>>
Loan offer	3	4	2	2	
Conference invitation	1	1	0	0	
Reminder of the previous message	0	0	6	8	
Offer friendship, erotica	0	0	2	2	
Unblocking a bank account; actualize account	0	0	2	2	
Unspecified message, call to communicate	0	0	4	5	

The trend is characterized by the symbols (used in Tab. 2, 3, and 4):

- == equal,
- = nearly equal (<2% difference),
- > slightly greater (2-5% difference),
- >> greater (>5% difference),
- < slightly smaller (2-5% difference),
- << smaller (>5% difference).

The content of the message is important for recognizing phishing emails (see Tab. 3). The content of the message was categorized into five groups: BUSINESS, CHARITY, FUND, TRANSFER, and OTHER. Another important element for recognizing phishing email is impersonal greeting. This factor appeared in all emails (see Tab. 4).

Table 4: Impersonal greeting in the phishing emails

Greetings in the email	2020/11		2021/05		TREND
	Sum	%	Sum	%	
NO greetings	26	34	17	21	<<
Hello (Beloved, Dear, Friend, Sir/Madam)	10	12	16	19	>>
Dear or Dearest (Beloved, Beneficiary, Friend, in Christ, Sir/Madam)	26	34	20	24	<<
Good Day (Sir/Madam)	3	4	7	8	>
Attention	2	3	2	2	=
Hi, Hey, Ahoj	2	3	5	6	>
May Allah Bless You; Assalam Alikum; May the peace of God be with you	0	0	2	2	>
Business (Partner, Customer, Nominee, Empfänger, Client)	0	0	6	7	>>
Greetings, Gruesse	8	10	8	10	==
Congratulations	0	0	1	1	=

Analysis from the sender point of view includes two parts: Company/Organization and Person. In the first experiment were 24 company/organization senders; of which four were used repeatedly (two times Bank of America New York and Bill Gate Foundation, three times Exxon Mobil and The United Nations). In the second experiment were 33 company/organization senders; of which five were used repeatedly (two times Bin Omran

Trading and Contracting Company, Exxon Mobil, Regional Bank of MSB Limited, The United Nations, and The World Bank). In both experiments were used repeatedly three companies/organizations: Bank of America New York, Exxon Mobil, and The United Nations.

From the Person point of view were in 76 phishing mails mentioned in the first experiment 43 male and 29 female senders; of which repeatedly three were used (Jean Phillip, Melinda Cruz, and Peter Alexandra); in four emails was no name. In the second experiment were in 84 phishing mails mentioned 54 male and 22 female senders; of which repeatedly four were used (Caroline Rabin, Grace Edward, Paulux Kasun, and Peter Alexandra); in eight emails was no name. In both experiments was used repeatedly only one name Peter Alexandra.

In addition to the above listed characteristics of phishing emails and statistics in Tab. 2–4, it is possible to analyze, the other characteristics, for example, the country of sender, email account, etc.). Another popular target for phishing email analysis is the text of the subject line of the message. It can therefore be stated that it was not possible to establish any common trend in all analyzed phishing emails for both experiments. There were changes in individual parameters, so more experiments will be needed.

4. Research of teaching and testing systems

Before creating our own education and training system in the field of cyber security, it is first necessary to examine existing solutions and use the knowledge gained. One of the most promising training systems is the Kaspersky ASAP (Automated Security Awareness Platform) (Kaspersky ASAP, 2021). It is based on the assumption that human behavior causes most breaches of corporate data, so raising employees' awareness about cyber security is one of the best ways to take action against cybercrime.

The structure of the ASAP program includes Pre-lesson test and interactive lessons (max. 10 min.), Summary and repetition of knowledge, Examples from life and testing, and Simulated phishing attack. The ASAP system is commercial, complex and, therefore, expensive. It is not easy that system to obtain for the needs of the university, and for the University of Defense, the product of a Russian company can pose a security threat. Although we could not observe ASAP in practice, it is inspiring for the design of our own teaching system.

Another training method is a gamified phishing training product (HOXHUNT, 2021), which is based on “real-life” training. This simple add-on is installed on your preferred email client, which serves to report phishing emails that are regularly sent to your email inbox. Your reaction to these emails is evaluated and you are regularly informed of your ability to resist email fraud.

The thing we liked most about this method is the so-called micro-training which informs the user about key indicators that could be found in the last received emails. This system is great, but it is only available as a paid service. Also, it completely relies on the function of an additional utility that is not normally found in email clients and thus teaches a habit that would not normally exist (see Fig. 2).

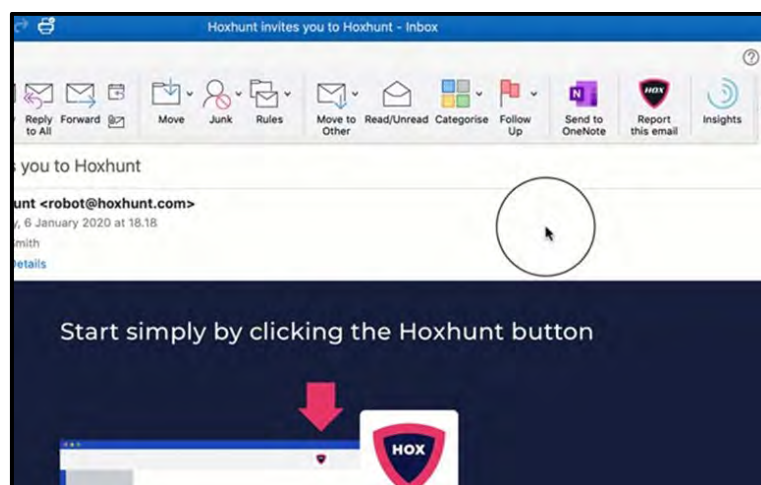


Figure 2: Hoxhunt add-on button (HOXHUNT, 2021)

A simple Slovak phishing test on the website of the Computer Incident Resolution Unit (CSIRT.SK, 2021) contains 17 questions; see the user interface in Fig. 3. The answers in the test are not limited in time and any incorrect answer is immediately indicated. Email messages are displayed sequentially, and the test participant must decide whether the message is forged or correct. At the end of the test is the evaluation and links to further study. The test is definitely useful, but its problem is that the vast majority of emails are in Slovak, while our research shows that nearly 90% phishing emails are in English, see Tab. 2.

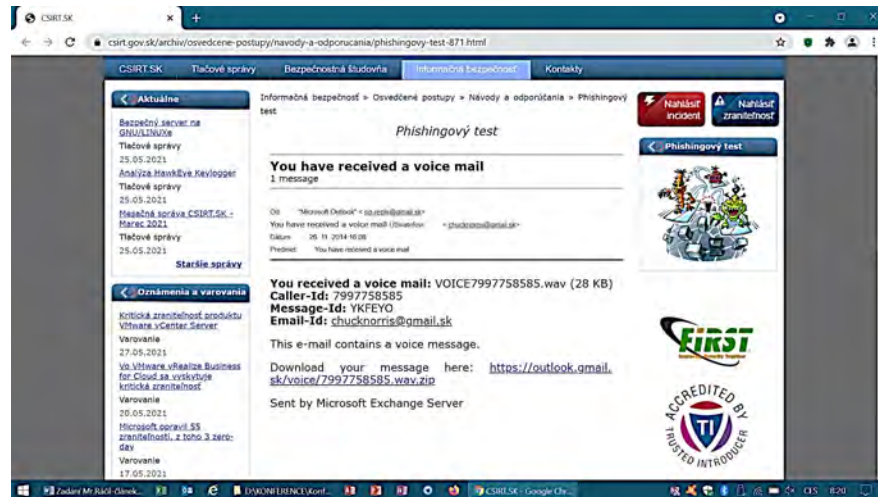


Figure 3: User interface of the Slovak phishing test (CSIRT.SK, 2021)

Examples of free training platforms are phishing tests on site hook security (HOOK SECURITY, 2021) and phishing box (Phishing box, 2021). Training involves showing you an image of the received email with the sender's address and then letting you decide if the email is phishing or not. The test has roughly 10 emails, and it seems that on each run the same email in the same order is shown. In the end, you are informed of your results. There is no explanation for your mistakes or why each email is considered phishing, so you may only guess the reasons. This test reflects the sad reality of free fraudulent training solutions.

All of the previously tested applications are not comprehensive and accessible solutions that could improve overall awareness of how to successfully counter not only mass phishing campaigns, but also targeted attacks on specific companies or individuals. In the following section, an overall solution for an educational and training system that fulfils these ambitions will be presented.

5. Education and training

The concept of our proposed system will be introduced in this chapter. The main goal is to create a new publicly available training system that will help improve the ability of users to distinguish phishing emails from harmless ones. The prerequisite for such a system is not only affordability but also scalability of questions according to the target students, where students will get complete feedback immediately. The system must be user-friendly and adaptable to the environment for which it is designed.

The key element we will focus on in our design is the ease of use, both for the user and the operator. On the operator side, it is necessary to ensure minimal personnel and hardware complexity, while for the user the simplest possible access to skill training to recognize fraudulent emails must be ensured. These goals are by no means new or innovative, as can be seen in the previous section where we have reviewed the existing systems, but we strongly believe that the combination of our method will bring something new.

We will first look at the system from the user's point of view (see Fig. 4). Of course, it is necessary to register on the system, but complicated forms can often discourage a potential user, as we found out in our research of existing systems, so we decided to choose a simple process where the user enters only his email address and agrees to the terms of use. Then they have to confirm the registration by clicking on a link emailed to them.

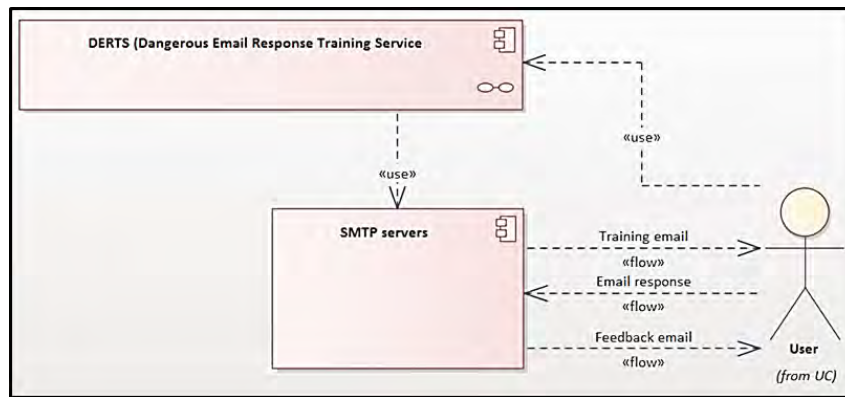


Figure 4: Simple component view

Now to the very method used to educate and train users. The method we have chosen is based on our experience as teachers—but also on common knowledge—that we learn best from our own mistakes. Our system will therefore simulate attacks on persons who register, similarly to the Hoxhunt method. From the user's point of view, nothing much will change after registering on our system, only that they will start receiving a little more fraudulent emails than usual. These emails will be generated by our system. Of course, the evaluating reaction of a user is the most important part of the whole process, but in contrast to the Hoxhunt method our system will implement no email client add-on.

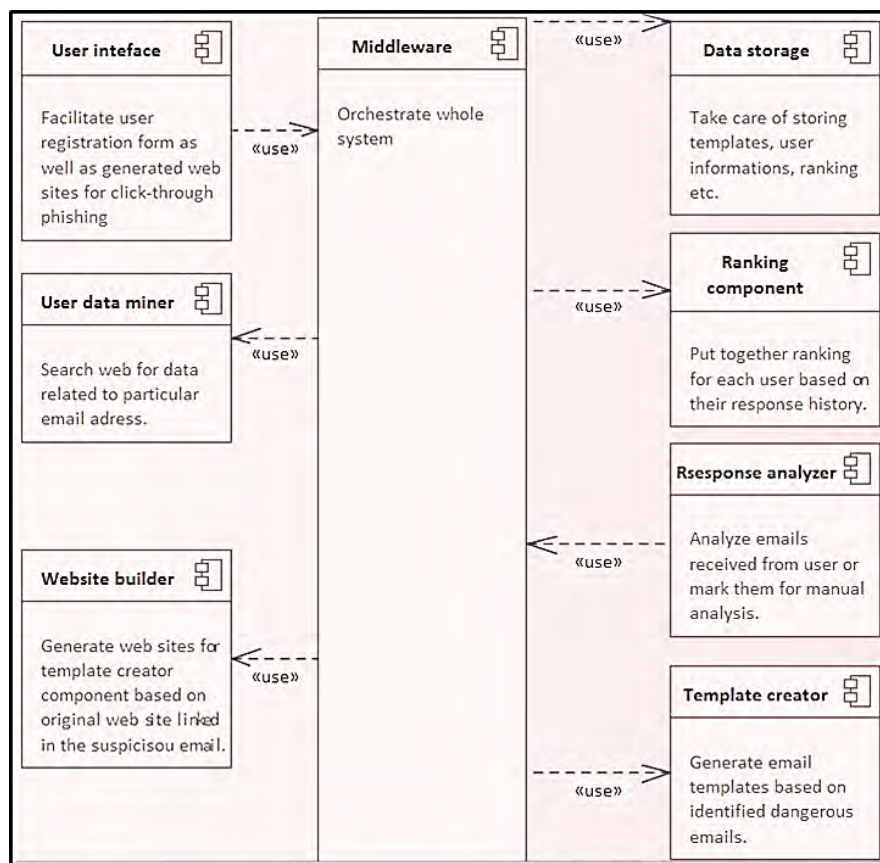


Figure 5: Component inside DERTS

If the user does not recognize the dangerous email or reacts badly to the email, they will immediately receive feedback in the form of an email message containing all the key information about this incident. This element is the cornerstone of our training and education system. Immediate feedback has a significant effect on whether knowledge is permanently acquired, and thus significantly increases the resilience of the person in the case of similar attacks [https://link.springer.com/article/10.1007/BF03395471]. The content of the feedback is of utmost importance. We have identified two key elements that should be included in the response. The first is information about what impact this mistake could have on the user and their surroundings. The second is

information about the error, and how to proceed correctly next time. So far, the system has been discussed only from the user's point of view, but now we will explain some key principles that will be hidden from the user. First of all, I stated at the beginning that our system does not require filling in of any complicated forms; this, besides facilitating the registration process, has practical purposes. An attacker often has no personal information except an email address, and must find the necessary information or derive it from the email address. The system will have a special module that takes the address provided by the user and tries to find additional information about the person. This information is then stored with the person along with the sources (so that the user becomes aware of it). If the system manages to find this information, the user will be informed of this fact. If this information is not found, the system will work as if it did not know it. This means trying to retrieve them from the user using fraudulent emails. If system fails to do so it will use types of fraudulent emails that do not require knowing those user's information (see Fig. 5 and 6).

A key element of the system is a module that will monitor the success of users in recognizing suspicious emails. The module will monitor how a user reacts when they receive a fraudulent message, store information about the response, and, based on the data obtained, compile an assessment of the user's resilience. This evaluation has two purposes. First, it will serve as a basis for regular evaluations. Second, it will serve to regulate the number and types of suspicious messages sent to a particular user, by creating a rating system. In other words, if a user proves resistant to a particular type of phishing, they will start receiving fewer of these email messages and ones that are far more sophisticated. The result should be an effective model of education, targeting only weak areas of individuals and reducing learning time.

The heart of the whole system will be the email template creator. This module will have the task of automatically compiling email templates from collected suspicious messages. We currently have a small database of suspicious emails; some of these emails have already been analyzed in the previous section of this article. An algorithm for creating a database of patterns from existing messages will be run on these emails. Subsequently, this pattern of database could be expanded with emails forwarded by users for check. This will be handled in the user response analysis mode. In case this module is undecided about the harmfulness or otherwise of a message, its classification will be left to system operators. Of course, it will also be possible to add patterns to this database manually. The template creator will also classify patterns and difficulty levels and, for some messages, artificially insert fraudulent message identifiers to ensure appropriate difficulty scaling.

The creation of the template module is entirely dependent on our research into phishing emails. The analysis of phishing emails, which is addressed in chapter three of this paper, provided us with the basis for creating templates for educational emails. It is the ability to categorize phishing emails that is quite crucial to building a quality educational system. This is because it is essential that we are able to break down the entire large and complex issue of phishing emails into smaller blocks (categories). The abilities of the users that our system will educate will be tracked across the categories to ensure that the educational process is optimized. The reason for focusing on each category is that each person has a different resistance to different tactics of email scammers, some will respond to emails offering a job, others to an investment opportunity. Finding patterns in the scam emails was also quite crucial to ensure that the templates created, and their distribution were as credible as possible.

A component closely related to the template creator is the module for generating feedback. This module generates feedback on individual patterns based on identifiers and possible reactions. It works closely with the user response analysis module. The feedback itself, as mentioned earlier, will contain information about the user's error, depending on their response, as well as a set of information about all the identifiers in a particular message, which should have warned the user that the email message is dangerous.

The task of the user response analysis module is to evaluate a possible response. It will first identify if the analyzed email was sent by the system. If so, it will just send order sending of feedback to the user. If the email is from an external source, the module will try to match it with an existing template. If it cannot find a match for the email, it will send it to the template creator to implement this type of attack on the system for future use.

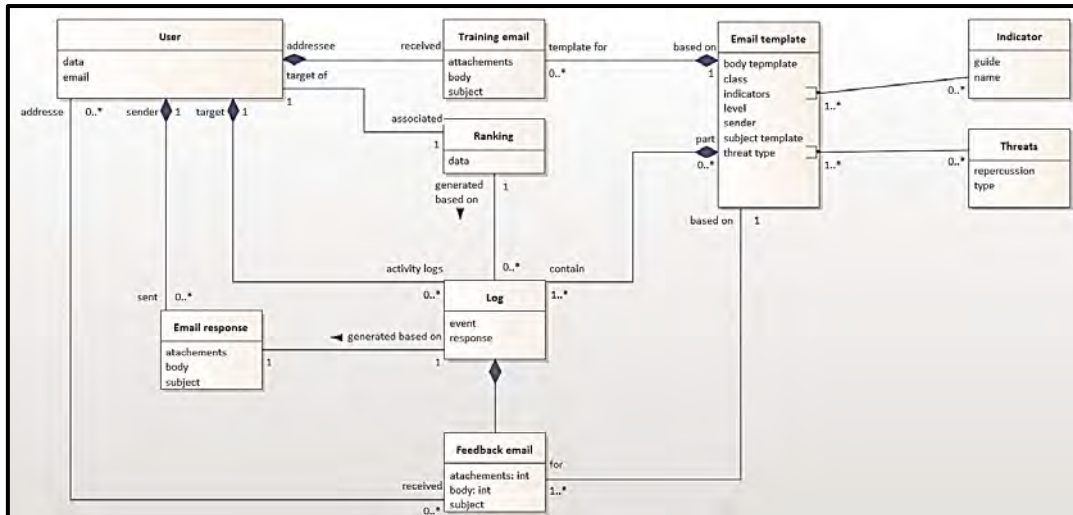


Figure 6: Domain model of our proposed system

An element not mentioned yet is a web service that will host various fraudulent forms, in order to enable the most accurate simulation of the attack and monitoring of user behavior. This web service will serve as a support tool for various types of emails that use click-through links. Links to these sites will be inserted into specific templates. Website creation will be automated from the received or manually inserted dangerous emails with click-through links as part of the template creation process. Last but not least, we plan to include a fabricated “malicious” code in different kinds of attachments as part of email message training. This will be used to simulate similar real-life attacks and report on users’ opening of dangerous attachments.

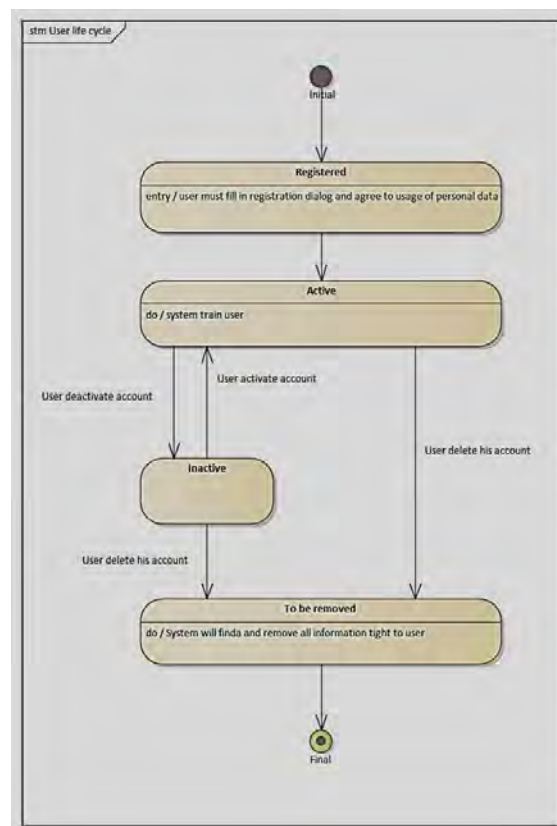


Figure 7: User lifecycle diagram

It is worth emphasizing that we do not plan to collect information about users in our system other than their email addresses, data obtained from widely available sources (name, surname, etc., which, nevertheless, can be removed on request), and historical success data. No email messages, sent files, and information entered into

forms will be stored in the system, thanks to the design of the template and evaluation modules. We also plan to develop the system as open-source, because our goal is to increase general security on the Internet.

To better explain the functionality of the proposed system on a basic level we created new figures. Fig. 7 shows the user lifecycle in an UML state diagram. As was already stated the system was designed to be as simple and easy to use as possible, so the only thing user needs to do is signup and this starts his training process.

Fig. 8 shows one instance of the training process in an UML activity diagram. This process will repeat itself with a frequency that will be based on user training progress more experienced users will have to train less frequently. On the left part, we can see gathering information necessary for creating training email in the middle is evaluation of user reaction to send email and on the right, we record information to the database to be used in the next round of training for the particular user.

6. Conclusions

The article is focused on the issue of phishing emails. Phishing emails have been analysed in order to recognize and understand the threats they pose. If phishing threats must be successfully guarded against, there is need to be able to recognize phishing emails. Teaching and testing systems are used for this purpose. After exploring some of the existing ones, a new teaching and testing system for the university is created. It is in the design stage and a subsequent stage will be devoted to its development.

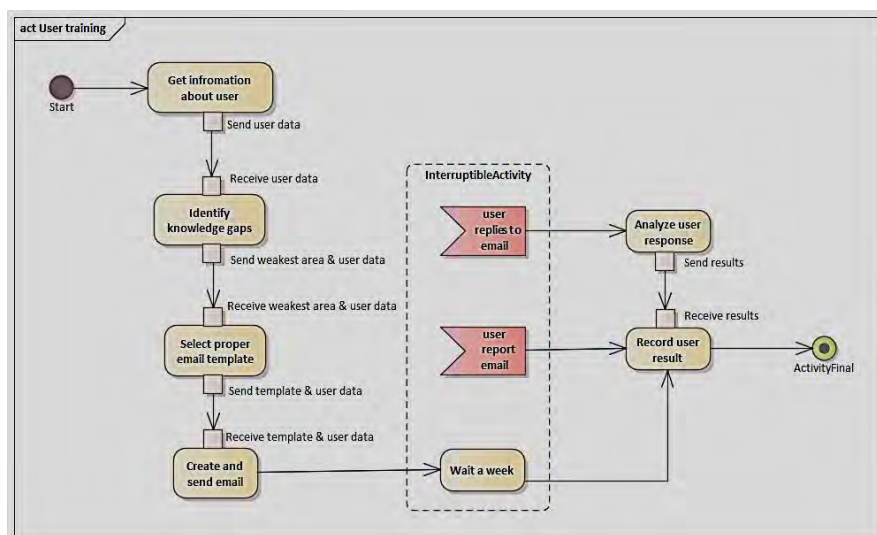


Figure 8: Training cycle diagram

References

- Butavicius, M., Parsons, K., Pattinson, M., and McCormac, A. (2015). "Breaching the human firewall: Social engineering in phishing and spear-phishing emails". Proceedings of the 26th Australasian Conference on Information Systems.
- CSIRT.SK. (2021). "The phishing test". [online], <https://www.csirt.gov.sk/osvedcene-postupy/navody-a-odporucania/phishingovy-test-871.html>
- Dhamija, R., Tygar J. D, and Hearst, M. (2006). "Why phishing works". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). Grinter, R., Rodden, and all. Eds. CHI '06. ACM Press, New York, NY, 581-590. DOI= <http://doi.acm.org/10.1145/1124772.1124861>.
- Dodge, Jr. R.C., Carver, C., and Ferguson, A.J. (2007). „Phishing for user security awareness“. Computers and Security, 26(1), pp. 73 – 80.
- HOOK SECURITY (2021). "Phishing test". [online], <https://www.phishingbox.com/phishing-test>
- HOXHUNT (2021). "Phishing awareness training that employees love". [online], <https://www.hoxhunt.com/gamified-phishing-training-platform>
- Kaspersky ASAP (2021). "Automated Security Awareness Platform", Webinar Computerworld on March 24, 2021.
- Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., and Hong, J. (2010). "Teaching Johnny Not to Fall for Phish". ACM Transactions on Internet Technology, 10(2), 7. [online], https://www.researchgate.net/profile/Steve-Sheng/publication/220169843_Teaching_Johnny_not_to_fall_for_phish/links/00463531e0f42496d6000000/Teaching-Johnny-not-to-fall-for-phish.pdf?origin=publication_detail

- Liu, Z., Zhou, L., and Zhang, D. (2020) "Effects of demographic factors on phishing victimization in the workplace". Proceedings of the 24th Pacific Asia Conference on Information Systems: Information Systems (IS) for the Future. [online], <https://aisel.aisnet.org/pacis2020/75/>
- Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., and Jerram, C. (2013). "Phishing for the truth: A scenario-based experiment of users' behavioural response to emails". IFIP Advances in Information and Communication Technology, 405, pp. 366-378.
- Phishing box (2021). "Online Phishing Test". [online], <https://hooksecurity.co/phishing-test>
- Sumner, A. and Yuan, X. (2019). "Mitigating phishing attacks: An overview", Proceedings of the 2019 ACM Southeast Conference, pp. 72-77.
- Takata, T. and Ogura, K. (2019). "Confront Phishing Attacks - From a Perspective of Security Education". 2019 IEEE 10th Conference on Awareness Science and Technology, iCAST 2019 – Proceedings, Article number 8923444.

Don't Drink the Cyber: Extrapolating the Possibilities of Oldsmar's Water Treatment Cyberattack

James Cervini, Aviel Rubin and Lanier Watkins

The Johns Hopkins University, Baltimore, USA

jschaf12@jhu.edu

rubin@jhu.cs.edu

lanier.watkins@jhuapl.edu

Abstract: Water treatment represents an essential critical infrastructure sector which has a direct impact on the health and well-being of its customers. Water treatment is often performed by municipalities with very limited budgets for cybersecurity resources. These underfunded, high-impact, targets represent an emerging cyber warfare attack-surface paradigm which poses a direct threat to the quality of life for millions of people. On February 5th, 2021, a water treatment plant in Oldsmar, Florida was the victim of an attempted cyberattack. This attack commanded the system to add a dangerous amount of sodium hydroxide to water which supplied thousands. Direct exposure to sodium hydroxide causes painful burns to the exposed area with permanent internal damage likely upon ingestion. A system operator noticed this malicious behaviour and corrected the situation, minimizing the attack's impact. This paper outlines the attack and illustrates how minor modifications to the attacker's tactics, techniques, and procedures could have resulted in a cyber-derived catastrophe for thousands of unsuspecting citizens. Lastly, this paper explores the effectiveness of various low-cost cyber-physical security technologies when pitted against differing attacker models in these theoretical scenarios. These cybersecurity solutions are evaluated by cost, ease of use, implementation difficulty, and ability to support safe operation continuity when faced with adversary behaviour. The results of this evaluation illuminate a path forward for low-cost threat mitigation which increases the difficulty to compromise these critical cyber-physical systems. With attacks targeting industrial control systems on the rise, the Oldsmar water treatment cyberattack represents more than an individual incident, it can be viewed as a reflection of the current status of thousands of similar critical infrastructure systems that have yet to be caught in crosshairs of a competent and willing adversary with financial incentives and cyber warfare mission requirements serving as impetus for adversary willingness and any resulting large-scale cyber cataclysm.

Keywords: operational technology, water treatment, Oldsmar Florida, cybersecurity, cyber warfare, cyber attack

1. Introduction

The water and wastewater systems sector is labelled as a vital critical infrastructure sector by the United States' Department of Homeland Security (DHS), describing this sector as "essential to modern life" (DHS, 2015). Water treatment responsibilities are commonly performed by municipalities with meagre budgets for cybersecurity resources. According to a report which surveyed water sector utilities, 44.8 percent claimed less than 1 percent of their overall budget was dedicated to operational technology (OT) cybersecurity (ThreatLocker Inc., 2021). The high impact of water treatment disruption and malicious manipulation when paired with an underfunded defensive cyber-posture results in an emerging cyber warfare paradigm which poses a direct threat to the quality of life for millions of people. The 2021 water treatment cyberattack orchestrated in Oldsmar, Florida represents a relevant example of cybersecurity shortcomings prevalent in underfunded municipalities. While Oldsmar is home to approximately 15,000 residents, coordinated cyber warfare campaigns targeting these vulnerable critical systems represent asymmetric threats which could impact millions.

This paper contributes to the domain by researching a relevant OT municipal cyberattack with unrealized impact while framing it in the context of cyber warfare at scale. Furthermore, this paper illustrates how an increased likelihood for impact realization is possible while outlining mitigations mindful to the operational constraints of municipalities. Section two of this paper will describe and dissect this attack, with section three ultimately showing how slight modifications to the attacker's tactics, techniques, and procedures (TTPs) could have resulted in a devastating cyberattack. Subsequently, section four describes low-cost mitigations which aim to mitigate the exhibited threats with various mitigation types described in each sub-section. Lastly, section five concludes the paper with closing thoughts.

2. The Oldsmar Water Treatment Attack

The attack began on February 5th, 2021, at 0800 where the attacker leveraged compromised credentials and the common remote access software TeamViewer to login to a plant operator's console. Following the login, the attacker immediately logged off and disconnected the session (Pinellas County, 2021). This was likely the attacker confirming system, connection, and credential validity. At this point the attacker could have also taken

a screenshot of the console supporting the planning of further actions. The console was manned by an operator at this time and the anomalous behaviour was mentally noted, but no action was taken. Later that day, at 1330 the attacker logged back into the operator console and used the plant's human machine interface (HMI) to command the increase of sodium hydroxide, also known as lye, to water which supplied thousands of people. Lye exposure can result in painful burns and permanent damage if ingested. The lye was increased from 100 parts per million (ppm) to 11100 ppm, an increase of over 100 times the intended amount. The operator observed this anomalous behaviour and commanded levels back to normal values. The operator then notified a plant supervisor, and the TeamViewer access was removed. This attack also occurred in the days leading up to the Super Bowl on February 7th which was being hosted 12 miles away from the facility, likely drawing more people to the area. According to county officials, it would have taken 24 to 36 hours of an increased mixture setting to have the malicious chemical mixture reach a point of distribution where it could have been consumed by customers. Additionally, county officials cited the use of pH level sensors with alarm reporting functionalities which would have notified operators of an imbalance (Pinellas County, 2021).

This attack lacks much of the complexity seen in previous cyberattacks targeting OT. Current reporting suggests the attacker did not attempt to evade detection and the attacker did not deploy custom scripts or programs, instead opting to leverage the existing operator interface to command their effect. While some would see this as an indicator of a low-skilled adversary, it is often the case that an attacker, regardless of skill, will use the easiest and fastest means to achieve a desired outcome. This can result in highly skilled attackers using techniques seen as simplistic or effortless. Therefore, it can be difficult to assess the attacker's skill in this case with the only potential indicator being the assumption that they did not meet their desired effect. Table 1 below breaks down relevant characteristics of the water treatment attack that occurred on 02/05/2021.

Table 1: The characteristics of the Oldsmar water treatment cyberattack

<i>Attack Characteristic</i>	<i>Oldsmar Water Treatment Attacker</i>
Attack Date	02/05/2021
Attack Time	1330
Access	TeamViewer Remote Service
Attack Execution	Process HMI
Detection Evasion	None Reported
Persistence	None Reported
Response Function Obstruction	None Reported
Operational Impact	▪ Momentary Control Manipulation ▪ Reduction of System Confidence

This table incorporates the current understanding of the attack's attributes and maps them to characteristics some of which are derived from the MITRE ATT&CK for industrial control systems (ICS) (Alexander et al, 2020).

3. Modification of Attacker Tactics Techniques and Procedures

Using the 02/05/2021 attack as a baseline and modifying certain attack characteristics generates two new theoretical attacker models. These two new attacker models vary by operational preferences for stealth and persistence but share the same mission of impacting the delivery of potable water, resulting in either an inability to deliver potable water or the delivery of dangerous water. Attack success is defined by the ability to command the increased addition of lye for the 24-to-36-hour timeline provided by county officials and the ability to degrade or eliminate the situational awareness provided by the system's pH sensors (Pinellas County, 2021). Given its proven success, all three attacker models will share a commonality of access via the TeamViewer remote service. A high-impact differentiator between the theoretical attacks and the 02/05/2021 attack is the date and time of attack execution. This temporal differentiator requires no additional skills to utilize and results in a delayed response and higher chance for mission success. The Oldsmar water treatment plant's normal hours of operation are publicly available via open-source information (City of Oldsmar: Water Division, 2021). Leveraging this information to launch attacks outside the normal hours of operation, both theoretical attackers will execute their operations just eleven hours and thirty minutes later compared to the 02/05/2021 attack, opting for an attack on 02/06/2021 at 0100 hours. Executing the attack at this time reduces the likelihood of an operator noticing the remote access and could increase the response time for lye level correction. For the sake of realism, neither theoretical attacker will leverage supply chain compromises, zero-day exploits, or low-level firmware manipulation as these concepts require a substantial amount of time and resources to employ. Table 2 below displays the various attacks and their TTPs.

Table 2: A comparison of attacker model characteristics.

<i>Attack Characteristic</i>	<i>Oldsmar Water Treatment Attacker</i>	<i>Theoretical Attacker A</i>	<i>Theoretical Attacker B</i>
Attack Date	02/05/2021	02/06/2021	02/06/2021
Attack Time	1330	0100	0100
Access	TeamViewer Remote Service	TeamViewer Remote Service	TeamViewer Remote Service
Attack Execution	Process HMI	Process HMI	Malicious Background Process
Detection Evasion	None Reported	Ransomware	System State Spoofing
Persistence	None Reported	None	Execution on Start-up
Response Function Obstruction	None Reported	Ransomware	▪ Command Blocking ▪ Alarm Suppression
Operational Impact	▪ Momentary Control Manipulation ▪ Reduction of System Confidence	▪ Control Manipulation ▪ Reduction of System Confidence ▪ Data Destruction ▪ Lengthy Inability to Deliver Potable water ▪ Possible Delivery of Unsafe Water	▪ Lengthy Control Manipulation ▪ Delivery of Unsafe Water

3.1 Theoretical Attacker A

Theoretical attacker A is impartial to attack discovery, instead opting to quickly impact the system while creating a difficult path for system recovery. Following the 0100 command to increase the lye amounts in the water supply, theoretical attacker A launches a ransomware attack on the system's computers to include operator stations. This ransomware attack reduces the operator's ability to control the system and can stifle the alerting of the pH imbalance due to a lack of situational awareness provided by their system interfaces. Communication could also be impeded if ransomware spreads to e-mail servers. A ransom note requesting bitcoin in exchange for decryption would manipulate responders into thinking the attack was financially motivated and concealing the attack's true intentions. Ransomware also has the added impact of encrypting log files, rendering them useless and increasing the difficulty of attack attribution.

According to a recent study, the number of reported ransomware attacks continue to increase yearly, one of which was the colonial pipeline ransomware attack (Reeder et al, 2021). This example illuminates the reality that malicious actors are targeting critical infrastructure systems which provide key services. This relevant example and continuing trend of ransomware use supports the realism of this theoretical attack scenario. Another emerging trend in ransomware is what is called "double-extortion" ransomware where attackers will apply additional pressure by threatening the release of sensitive stolen data in addition to the encryption of systems (Payne et al, 2021). To this end, it is possible that threats of malicious physical process manipulation carried out by dormant logic bombs will accompany future ransomware attacks on critical infrastructure to add additional pressure to the ransom. To illustrate the timelines associated with ransomware, a ransomware attack on the city of Baltimore took over a month to recover (Petcu, 2020). Acknowledging the differences in scale between the two systems, a report polling 2690 information technology (IT) professionals claimed that 66 percent did not believe their organization could recover from an unpaid ransomware attack in less than five days (Veritas Technology, 2020). Therefore, it is generally unlikely that full system functionality would be restored in the 24-to-36-hour timeline. Additionally, the odd hours of the attack delays initial attack discovery and subsequently the recovery process. Lastly, the unseen lye command and potential degradation of the system's pH sensing due to encrypted operator interfaces enables the potential delivery of unsafe water. In the event plant personnel discover the lye manipulation but are unable to access encrypted interfaces to assert positive control, a possible response could be to halt the delivery of water until systems are reconstituted. In both scenarios, the attacker has achieved their mission and impacted the quality of life for thousands.

3.2 Theoretical Attacker B

Theoretical attack B requires additional planning and domain understanding to execute but ensures guaranteed mission success. Attacker B relies on stealth to ensure that operators are unaware of the true state of their systems until dangerous water reaches customers. This attack requires a moderate understanding of cyber-physical systems and cyberattacks. This attacker gains initial system access via the TeamViewer remote access

application. Instead of continuing to use TeamViewer, attacker B establishes a malicious background process which provides subtle backdoor access unobservable to an operator under normal operating conditions. This process is configured to run at start-up, granting it persistence upon reboot cycles. This gives the attacker a long-term foothold to gain additional information on the system's processes, devices, and configuration. With this newfound system understanding, the attacker chooses to execute a man-in-the-middle (MITM) attack. This attack commands the system to add dangerous levels of lye to the water while reporting normal values to the operator. This can be accomplished in multiple ways and the attacker chooses whichever presents the highest chance of success based on initial reconnaissance.

In this theoretical scenario, the attacker chooses to modify the programmable logic controller's (PLC) ladder logic responsible for lye addition and pH reporting. PLCs are ruggedized real-time controllers prominent in many ICS environments which execute pre-programmed logic and interface with physical systems via control signals to actuators and sensors. The attacker reprograms the device, hard coding a dangerously high value of lye to be added which would otherwise be controlled by the operator or the device's original logic. The attacker also modifies the scaling of any reported values related to lye introduction and pH sensing to reflect normal values. This strips away awareness of the system's true status provided by the physical sensors connected to the PLC. An additional benefit that stealth provides is the inability to discern cyberattack from system failure. In the event plant personnel were able to detect the pH imbalance through an out-of-band channel, they would more likely attribute the anomaly to a faulty sensor or hardware failure then to a cyberattack. As a result, any hardware replacements would prove ineffective considering the attacker's foothold on the system. This attack uses long-term access and stealth to ensure the lye value's change lasts at least the 24 to 36 hours specified by county officials. It also considers the system's pH sensing and circumvents this via a MITM attack. Even if these operational requirements were not outright stated, the attacker's long-term foothold would have provided sufficient time to understand the system process and form attack requirements. Given this attacker's stealthy approach, it is highly likely that the first indications operators would receive of this attack would be calls associated with consumption of dangerous water, resulting in a successful mission for the attacker.

4. Theoretical Application of Potential Low-Cost Mitigations

The following subsections outline various solutions which attempt to mitigate the threats shown throughout this paper. These solutions prioritize threat mitigation while adhering to the budgetary and personnel constraints limiting many municipal utilities. Even so, all these solutions are directly dependent on the understanding of the current system and a lack of detailed system and configuration knowledge could reduce the impact of these mitigations. Additionally, any misconfiguration of these capabilities could result in a less effective mitigation solution. Lastly, these solutions are best utilized in combination with policies, procedures, and training that outline capability use.

4.1 Remote Access Limitations

The first mitigation is to establish and ensure a strong cyber-perimeter, this includes restricting and eliminating remote access wherever possible within operational parameters. Ensure devices are patched, prioritizing any internet facing devices. If remote access to plant data is required for situational awareness but positive control can be limited to on-site personnel, plants should consider implementing a one-way data-diode. If remote access is required for periodic remote maintenance, leave properly configured remote access systems offline, only bringing them online during coordinated maintenance time-windows to minimize internet exposure. Ensuring proper configuration and patching systems does not incur costs associated with new hardware, instead requiring minimal personnel time making it a cost-effective strategy to mitigate potential threats. One-way data diodes do incur hardware costs but are generally available for less than \$10,000 and significantly reduce risk if they are the only operationally relevant solution for remote access removal at a given site. Proper remote access limitations would have mitigated the 02/05/2021 attacker and both theoretical attackers. Even so, while ensuring a strong cyber-perimeter is the first line of defence towards threat mitigation, it must be assumed that an attacker will still be able to gain access and that mitigation measures within the cyber-perimeter should be implemented. The following sections outline solutions which apply to internal systems and aim to mitigate threats within these boundaries.

4.2 Anomaly Detection and Prevention

A heightened awareness of the communications traversing the internal network can illuminate malicious actors on a network, aid in attack response, motive discovery, and attribution while assisting in the differentiation between hardware failure and cyberattack. While the 02/05/2021 attack was blatantly observed, the two

theoretical attacks are executed such that potential observation is minimized. In this case, properly configured network security monitoring and logging would have contained the artifacts and alerting necessary to begin responding to the attack. However, the effectiveness of this mitigation relies on how operationally relevant its detection mechanisms are as anomaly detection only designed for traditional IT system threats lack much of the domain-specific context present in OT systems. For example, an anomaly detection system comprised entirely of IT threat signatures would be unable to discern the lye increase as malicious. Therefore, a low-cost OT-relevant solution is required to maximize situational awareness in the event of a cyberattack.

Security Onion and ROCK NSM are two popular and free open-source solutions to network monitoring (Burks, 2012) (RockNSM, 2019). While these tools are IT-centric, additional rules can be added and customized to provide more OT-relevance. Digital Bond's Quickdraw is an example of free open-source rulesets which can be added to alert on OT threats (Peterson, 2009). Additionally, Malcolm is a free open-source solution which has additional support for the communication types prevalent within ICS environments (CISA, 2021). The most effective anomaly detection stems from rulesets generated by an understanding of process knowledge. For example, if the command to increase the lye amount should never exceed a certain value, a specific rule can be written to alert on this occurrence. These customized operational-specific rules require a deep understanding of the OT system and process paired with an understanding of the IT infrastructure required to implement them. This network monitoring solution requires minimal hardware investment, only a host machine to run the software and a configuration of network infrastructure to support port mirroring which provides data for ingestion. However, this solution does require a significant amount of personnel time to configure, install, and monitor. OT-tailored solutions which utilize artificial intelligence to automatically baseline a system and alert on deviations of standard behaviour exist, but they require a significant capital investment compared to the free open-source solutions but require far less personnel investment. Therefore, a solution choice should be determined based on a site's available resources. However, regardless of the solution, policies and procedures which outline actions following alert generation should be implemented to best utilize the capability.

In the event theoretical attack A is successful, and operators arrive to find machines presenting ransom notes requesting payment for system decryption, examination of the anomaly detection artifacts highlights indicators of when and how the attack occurred. If the anomaly detection has enough OT specificity, responders would be alerted of the lye change and the attacker's true motivations. If the alerting is configured to be sent to offsite personnel via e-mail or short message service (SMS) transmission the attack response becomes timelier, minimizing the impact of both theoretical attacks' off-hour execution. Lastly, due to the anomaly detection system's ingestion of mirrored traffic, it can be segmented from the network such that a ransomware attack would be unable to modify its artifacts.

Theoretical attacker B's stealth makes detection more difficult, but not impossible. Up-to-date and customized OT rulesets would detect the reprogramming of the PLC allowing plant personnel to attribute system abnormalities to a cyber incident. Additionally, the associated alerts would attribute the attack to a source device, allowing attack responders to segment and investigate the system with the malicious background process.

4.2.1 Bump-In-The-Wire Solutions

Anomaly detection and prevention can be backfitted into existing systems using bump-in-the-wire (BITW) solutions. These solutions are implemented using hardware which sits on the network in between devices. They monitor each communication packet traversing the network, alerting and blocking traffic which trigger pre-programmed security rulesets. Unlike the network monitoring solutions showcased in section 4.2 which used port mirroring to ingest data, the BITW solutions are positioned in the network to be able to act when a rule is triggered as opposed to just alerting an operator. These solutions require an investment in hardware to place throughout the system. The size of an installation will determine the significance of the hardware investment. However, if financial resources are limited, the installation of BITW solutions can be prioritized to critical systems.

In theoretical attack A, the command to increase the lye amount could be detected and automatically dropped. Accompanying alerts would give responders a timeline of events and provide insight into the attacker's true intentions. A similar ruleset prohibiting device reprogramming unless expressly allowed through operator coordination would eliminate theoretical attacker B's MITM attack, alerting reduces the adversary's stealth advantages, and provide responders with awareness to support further remediation.

4.3 Security Orchestration

Security orchestration automates many of the actions a network defender would take in combating a cyberattack via pre-defined playbooks. Orchestration could be a huge benefit to installations with limited personnel or no personnel dedicated to actively defending the network. Free orchestration solutions allow for a pre-defined limit of automated actions to occur within a time window, but these constraints would likely be within most requirements of a small municipal control system. Orchestration itself requires a minimal investment of capital and personnel resources, but it does require a system with security solutions implemented to be most effective. For example, orchestration could automate the generation and implementation of a new firewall ruleset based on observed malicious behaviour. This scenario would require a competent network monitoring solution to observe and alert on the malicious behaviour and would also require a firewall compatible with your orchestrator to ingest the new ruleset. If the orchestrator does not have sufficient data to inform its decision-making or the proper mechanisms to enact impactful defences its effectiveness is suppressed.

The ability to automate the response and adaptation to cyber threats in real-time is a valuable asset and can negate the benefits of both theoretical attacks' off-hour executions. Assuming a compatible security architecture is in place, the automation of the response process helps to ensure recovery occurs within the 24-to-36-hour timeline, greatly reducing the likelihood for attack success.

4.4 Ransomware Mitigation

With ransomware's increasing prevalence in recent years, solutions specifically tailored to minimize ransomware's impact should be considered. Offline backups assist with system reconstitution in the event of a ransomware attack. Due to the generally static nature of ICS environment configurations, storing offline backups of OT systems should be performed. As an example, the configurations associated with HMI displays and the logic files of the PLCs go unchanged for long periods of time. Offline backups of these systems are crucial in the event your systems are tampered with, or reprogrammed. The infrequent back-up cycle of OT components results in minimal personnel time investment and the likely small amount of data results in inexpensive offline storage requirements. The system's IT components should be backed-up more often if possible. Even so, ransomware recovery methods for the IT components which automatically backup systems on an unwritable opal drive would quickly speedup recovery time from weeks to hours (Challener et al, 2018). The capacity to greatly minimize the impact of theoretical attacker A's ransomware or theoretical attacker B's PLC reprogramming would allow operators to quickly reassert positive control and situational awareness. With positive control re-established, the ICS levels would be discovered and corrected.

4.5 Training and Awareness

Training which attempts to increase the cyber-hygiene of personnel should be implemented to minimize unwitting insider threats. Periodic tests which help to understand the current state of cyber-hygiene within an organization should be regularly practiced. A common example of this being phishing e-mail tests. Additionally, personnel should be aware of the threat that cyber has on these systems and how they should respond given certain indicators or scenarios.

5. Conclusions

This paper described the 02/05/2021 attack on the Oldsmar water treatment plant and how it serves as an example for the current state of many systems responsible for providing basic needs. Theoretical scenarios showed how minor changes could have resulted in a dangerous situation for thousands of people. This paper's proposed mitigations increase the difficulty to execute these attacks while reducing their likelihood for success. The 2021-2022 Oldsmar annual budget contains only a single line directly relevant to this topic, "Continue network hardening measures around City's supervisory control and data acquisition at reverse osmosis water treatment plant and water reclamation facility." with no specifics describing how that will be accomplished or how success will be measured (City of Oldsmar, 2021). The shortcomings outlined in the paper are not isolated to Oldsmar and should be viewed as a general reflection of the municipal control system landscape. A landscape which could become a lucrative target for asymmetric cyber warfare operations. This paper hopes to be a motivator to augment existing systems with these mitigation concepts and serve as a potential influence for the requirements of the unbuilt systems of the future.

References

Alexander, O., Belisle, M. and Steele, J., 2020. *MITRE ATT&CK® for industrial control systems: Design and philosophy*, The MITRE Corporation: Bedford, MA, USA.

- Burks, D., 2012. *Security onion*.
- Challener, D.C., Kruus, P.S., Fink, R.A. and Farlow, J.F., Johns Hopkins University, 2018. *Apparatus and method for preventing access by malware to locally backed up data*, U.S. Patent 10,049,215.
- Cybersecurity and Infrastructure Security Agency (CISA), 2021, *Malcolm*, [online], <https://github.com/cisagov/Malcolm>
- City of Oldsmar, 2021. *City of Oldsmar Fiscal Year 2021/2022 Annual Budget*, p 111.
- City of Oldsmar: Water Division, 2021. *Hours*, Oldsmar, FL, USA.
- Department of Homeland Security, 2015. *Water and Wastewater Systems Sector-Specific Plan*, District of Columbia, USA.
- Heenan, R. and Moradpoor, N., 2016. *Introduction to security onion*. The First Post Graduate Cyber Security Symposium.
- Payne, B. and Mienie, E., 2021. *Multiple-Extortion Ransomware: The Case for Active Cyber Threat Intelligence*. In ECCWS 2021 20th European Conference on Cyber Warfare and Security. Academic Conferences Inter Ltd. p 331.
- Petcu A. G., 2020. *The Curious Case of the Baltimore Ransomware Attack: What You Need to Know*.
- Peterson, D., 2009. "Quickdraw: Generating security log events for legacy SCADA and control system devices", *Cybersecurity Applications & Technology Conference for Homeland Security* (pp. 227-229). IEEE.
- Pinellas County Sheriff's Office, 2021. *Treatment Plant Intrusion Press Conference*, Oldsmar, FL, USA
- Reeder, J.R. and Hall, C.T., 2021. *Cybersecurity's Pearl Harbor Moment: Lessons Learned from the Colonial Pipeline Ransomware Attack*.
- RockNSM, 2019. *What is ROCK*.
- ThreatLocker Incorporated, 2021. *Protecting Water Infrastructure Against Cyberattacks*, Maitland, FL, USA.
- Veritas Technologies LLC, 2020. *The 2020 Ransomware Resiliency Report*, Santa Clara, CA, USA.

An Ontology for Effective Security Incident Management

Sabarathinam Chockalingam¹ and Clara Maathuis²

¹Institute for Energy Technology, Halden, Norway.

²Open University of the Netherlands, Heerlen, The Netherlands.

Sabarathinam.Chockalingam@ife.no

clara.maathuis@ou.nl

Abstract: With the evolution of technologies like Internet of Things (IoT), there will be more and more connected devices in use around the world. This is one of the reasons why cyber security is critical to contemporary society as it makes the large majority susceptible to cyber-attacks. Such cyber-attacks not only impact confidentiality, integrity, and availability but also can cause physical damage. This is evident from cyber-attacks like Stuxnet and German steel mill. Effective security incident management plays an important role in minimising negative impact of such attacks mainly in terms of the organizations' finance, reputation, and personnel safety. Typically, the main phases of security incident management include: (i) preparation, (ii) mid-incident, and (iv) post-incident. There are diverse set of concepts like Structured Threat Information Expression (STIX) and Incident Object Description Exchange Format (IODEF) in the above-mentioned phases of security incident management. However, a comprehensive overview of different concepts and the relationships between such concepts in security incident management is missing. In this paper, we develop an ontology model with relevant concepts and their corresponding relationships between them especially in the mid-incident and post-incident phases of security incident management. Furthermore, we demonstrate the proposed ontology model using colonial pipeline example case study. The proposed model will help incident responders to operationalise concepts, by having a clear understanding on different concepts and their corresponding relationships, which in turn would also make the incident response more effective in practice.

Keywords: Cyber incident recovery, Cyber security, Incident response, Ontology, Security incident management

1. Introduction

There will be more and more connected devices in use around the world with the evolution of technologies like Internet of Things (IoT). Furthermore, with the emergence of remote working, people and critical systems in organizations are increasingly connected through the internet. At the same time, there is an increased rate of cyber-attacks during this Covid-19 pandemic (Lallie et al, 2021). Some of the well-known cyber-attacks include Brno university hospital cyber-attack (Pranggono et al, 2020), colonial pipeline ransomware attack (TechTarget, 2021) during this period. Such cyber-attacks may have different types of harms including physical or digital, economical, psychological, reputational, and societal (Agrafiotis et al, 2018).

Effective cyber security incident management plays a crucial role in reducing the dire consequences of cyber-attacks (Kulikova et al, 2012) especially in terms of the above-mentioned types of harms. There are four main phases of NIST incident response lifecycle: (i) preparation, (ii) detection and analysis, (iii) containment, eradication, and recovery, and (iv) post-incident activity. (Cichonski et al, 2012) also list relevant information/concepts corresponding to each phase of NIST incident response lifecycle. For instance, attack vectors, precursors, indicators, sources of precursors and indicators, incident analysis, incident documentation, and incident notification correspond to detection and analysis phase, whereas lessons learned, using collected incident data, and evidence retention corresponds to post-incident activity phase. Likewise, there are various globally accepted standards like Structured Threat Information Expression (STIX) and Incident Object Description Exchange Format (IODEF) in addition to concepts like contributory factors, and observations (test results) that are highlighted in other studies.

It is important for the incident responders to have an overview of relevant concepts which can aid an effective cyber security incident management in their organization. However, there is a lack of model that can provide a holistic overview of such concepts and their relationships. Ontologies have the capability to tackle this challenge especially based on their existing applications as it can help to represent, structure, and simulate the relevant concepts as classes and sub-classes in addition to labelling their relationships (Oltramari et al., 2015). Therefore, in this study, we will tackle the following research question (RQ): *"How can we develop an ontology model that*

provides a holistic overview of relevant concepts in security incident management?”. In order to address this RQ, we will consider the following research objectives (ROs):

RO1. To review scientific literature in security incident management and identify relevant concepts.

RO2. To develop and evaluate an ontology model which provides the holistic overview of the identified relevant concepts and their relationships.

The main contributions of this paper are as follows:

1. we identify relevant concepts in security incident management from scientific literature.
2. we develop an approach to provide an overview of the identified concepts and their relationships.
3. we demonstrate the proposed approach using an example case study.

The remainder of this paper is structured as follows: In Section 2, we describe the research method followed by the review of literature on security incident management especially covering the mid-incident and post-incident phases in Section 3. We propose the security incident management ontology based on the considered requirements followed by the demonstration through colonial pipeline example incident in Section 4. Section 5 highlights the related work on ontologies followed by conclusions and future work directions in Section 6.

2. Method

With the aim of designing a model that captures the security incident management process while focusing on the mid-incident and post-incident phases, this research intends to answer the above-mentioned RQ. To answer it, an extensive literature review followed by a case study were conducted in a Design Science Research approach (Peppers, Tuunanen and Niehaves, 2018; Hevner, March, and Park, 2004) using the Methontology methodology (Fernández-López, Gómez-Pérez and Juristo, 1997). Hence, the following activities are taken:

Specification: the objective, requirements, and the necessary knowledge for developing the model are defined.

Knowledge Acquisition: the information required for building the model is gathered from sources extracted during the literature review and the case study on an ICS cyber incident. An extensive literature review is conducted based on the steps considered by (Okoli, 2015; Cooper et al, 2018). The aim of this literature search is to identify relevant scientific literature in security incident management especially the mid-incident and post-incident phases.

Conceptualization: the information collected is conceptualized in taxonomies with concepts, characteristics, and relations between the concepts. The adoption of ontology for the proposed model is due to its richness, robustness, explainable nature, and its spread use in diverse decision-making processes (Mavroeidis and Bromander, 2017).

Formalization: the architecture of the model is built, is depicted in Figure 1, and described in Section 4.

Integration: other relevant models found during the review process are considered in related work. The model proposed is designed from scratch with important aspects found during this process.

Implementation: the software environment i.e., Protégé is prepared to develop the model using OWL (Ontology Web Language), which is a well-known approach used by researchers and practitioners in this field. Furthermore, the model is implemented capturing the concepts, their attributes, and the relations between these concepts.

Maintenance: the implemented model is updated to make sure that further operations are possible.

Evaluation: the evaluation of the model is done adopting a two-step process: (i) by considering its structure and consistency, and (ii) through exemplification based on a case study conducted on an ICS cyber security incident.

Documentation: the documentation of the model is done considering the descriptions provided during the development process. This documentation is embedded in different sections of this paper.

3. Review on Security Incident Management

(Lee et al, 2018) proposed a model for security incident handling based on the survey of existing incident handling models. The proposed model consists of three components including monitoring, incident response, and forensics. They highlighted the three key phases within incident response: (i) detection and diagnosis, (ii) incident analysis and response, (iii) post-incident analysis. Finally, they also listed the main actions corresponding to each phase. The actions corresponding to detection and diagnosis are to: identify suspicious behaviour, analyse precursor and indicators, look for correlating information, perform preliminary impact assessment and determine potential damage to prioritize incident, and categorize the event and classify the security level. (Staves et al, 2020) developed a framework that help operators in incident response and recovery operations. The proposed framework consists of four different phases including: (i) planning, (ii) preparation, (iii) mid-incident, (iv) post-incident. They also listed main actions corresponding to each phase. (Shinde and Kulkarni, 2021) reviewed different incident management approaches. Furthermore, Shinde et al. proposed an incident management framework with primary components including alerts, logs and data, existing knowledge, policies, frameworks, and tools. Similar to Lee et al. and Staves et al., Shinde et al. highlighted the different phases of incident management and the important actions corresponding to each phase. (He, Inglut and Luo, 2022) et al. proposed a malware incident response lifecycle based on the NIST incident response lifecycle. In this work, they also tackled the different phases of incident management and the main actions related to each phase.

Even though the phases highlighted in these works differ, it can be categorized into three different phases: pre-incident, mid-incident, and post-incident. The focus of this paper is on mid-incident and post-incident phase as there are tools/methods that can be applied across different organisations unlike the pre-incident phase. Additionally, as a part of relevant actions, the following themes are identified in these two phases for extensive literature review which includes root cause analysis, information sharing, forensics, and lessons learned.

Root Cause Analysis: (Findrik et al, 2017) proposed a high-level architecture based on Evidential Networks (ENs) approach to help operators during the detection and analysis phase of security incident management to determine the root cause of an incident. The key components of the proposed architecture include: (i) detectors, (ii) reasoning engine, (iii) containment engine, and (iv) remedial actions. Detectors play a similar role to Intrusion Detection System (IDS) which takes data (example: network traffic) as input and generate an event (example: alarm in case of malicious behavior). However, on a holistic viewpoint considering both cyber and physical data from different systems, to detect a cyber-attack. The reasoning engine based on ENs consider the events generated from detectors and determine the control system state as compromised, erroneous, or normal. This would support operators in understanding what they are dealing against and aid containment engine to choose appropriate containment actions from the list of containment actions. (Chockalingam et al, 2018) developed a framework which would help to develop Bayesian Network (BN) models for distinguishing attacks and technical failures. These models would help operators during the detection and analysis phase of security incident management to determine the root cause of an incident (cyber-attack or technical failure). There are three different types of variables in the developed framework which includes: (i) contributory factors, (ii) problem, (iii) observations (or test results). (Chockalingam et al, 2021) developed a BN model for the problem of incorrect sensor measurements in the flood management domain based on the proposed framework for distinguishing attacks and technical failures.

(Chockalingam et al, 2019) proposed a framework which would help to develop BN models for determining the failure cause in case the considered problem is caused by a technical failure or attack vector when the considered problem is caused by an attack. These models would support operators during the detection and analysis phase of security incident management to determine the root cause of an incident (failure cause in case of a technical failure or attack vector in case of an attack). (Piatkowska et al, 2020) proposed EN-based REASENS framework to support reasoning and determining the root causes of software rollout failures in the smart grid based on events collected from distributed and heterogenous sensors in a smart grid. The key components of REASENS framework includes: (i) sensors, (ii) causal reasoning engine, (iii) systems state (transients, misconfigurations, cyber-attacks). Sensors (or detectors) and causal reasoning engine are defined as in (Findrik et al, 2017).

Information Sharing: (Goodwin et al, 2015) proposed a framework for cyber security information sharing. Its building blocks include: (i) types of information exchanged (incidents, threats, vulnerabilities, mitigations, situational awareness, best practices, strategic analysis), (ii) actors involved in information sharing (Government, private critical infrastructure, business enterprises, IT companies, IT security firms, security researchers), (iii)

models of exchange (voluntary exchange models, mandatory disclosure models), (iv) mechanisms of exchange (person to person, machine to machine), (v) methods of exchange (formalized exchanges – based on agreement, security-clearance based exchanges, trust based exchanges, ad hoc exchanges), (vi) information formats (open response – any formats, unique information sharing, structured information sharing). (Alrimawi et al, 2018) developed a share incident knowledge approach across different cyber-physical systems (CPS) through two different meta-models including incident-patterns and CPS. The incident-pattern meta-model represents activity, scene (set of activities), crime script (set of scenes), asset, asset type (physical/digital), actor (offender/victim), resource (tool needed to perform an activity), type of resource (physical/digital), location, type of location (physical/digital), connection (between a location and other entities), and activity initiator. The CPS meta-model captures action, asset, asset type (physical/digital), containment, connection (physical/digital), actor, process (status: running/stopped), physical structure (part of the smart building layout), computing device (example: fire alarm). This helps to extract common incident patterns based on the above-mentioned meta-models and apply appropriate security measures in the future to prevent such incidents. (Rajamaki et al, 2019) highlighted technical standards for sharing cyber information including Structured Cyber Observable eXpression (CybOX), Threat Information eXpression (STIX), Trusted Automated eXchange of Indicator Information (TAXII). (Luijckx et al, 2015) also highlighted technical standards of sharing cyber information like CybOX, STIX, TAXII, and Malware Attribute Enumeration and Characterization (MAEC).

Cyber Incident Data and Analysis Working Group (CIDAWG, 2015) identified 16 data categories that can be shared anonymously to help organizations in security incident management especially by assessing their risks effectively, putting in place appropriate control measures. The data categories include: (i) type of incident (example: ransomware), (ii) severity of incident, (iii) use of information security standards and best practices, (iv) timeline (date of detection of a cyber incident in addition to date of effective control), (v) apparent goals (or motives – which could be financial), (vi) contributing causes, (vii) security control decay – a security control that is in place, but failed to withstand an incident, (viii) assets compromised/affected, (ix) type of impact(s), (x) incident detection techniques, (xi) incident response playbook – how an operator/decision maker respond to an incident (actions, methods, procedures and tools used), (xii) internal skill sufficiency, (xiii) mitigation/prevention measures, (xiv) costs, (xv) vendor incident support, (xvi) related events.

Forensics: (Van Vliet, Kechadi and Le-Khac, 2015) demonstrated a method to safeguard important information for digital forensic investigations in Industrial Control Systems using a wind turbine case study. They highlighted two important source of information which includes network and device data. The sources of network data include live network data (examples: raw network data, arp tables), historical network data (examples: host-based logs, firewall logs), other log files (examples: access point logs, historians). The sources of device data include running program data (examples: RAM dump, CHIP images), activity log files (examples: active processes, control room logs), transaction log files (examples: error logs, event logs). (Spyridopoulos, Tryfonas and May, 2013) investigated the state of SCADA operators to analyze cyber security incidents in addition to developing digital forensic readiness. They pointed the five different phases in a digital forensic process which includes: (i) examination, (ii) identification, (iii) collection, (iv) analysis, (v) documentation. They also described the information that should be obtained in each phase of the digital forensic process. For instance, during the examination phase of the digital forensic process, at minimum level, network diagrams, configuration details, change logs and authentication credentials needs to be obtained. Furthermore, in the identification phase, type of system, operating system used, type and manufacture of the PLCs, and network design and implementation need to be obtained. (Stirland et al, 2014) explores the challenges of forensic analysis in industrial control systems and proposes a forensic methodology for such systems. They highlighted the different phases in a digital forensic process which includes: (i) identification and preparation, (ii) identifying data sources, (iii) volatility assessment, containment impact analysis and preservation, prioritizing and collection, (iv) examination, (v) analysis, (vi) reporting and presentation, and (vii) reviewing results. This is like the phases highlighted by (Spyridopoulos, Tryfonas and May, 2013). (Spyridopoulos, Tryfonas and May, 2013) described the information that should be obtained in each phase of the digital forensic process in addition to example tools needed in each phase. For instance, example tools needed in the phase of volatility assessment, containment impact analysis and preservation, prioritizing and collection includes write blockers, firewire PCI card, HD camera, FTK imager, EnCase, Helix, TCP dump, data hashing tool, and text editor.

Lesson Learned: The literature that was identified as part of lessons learned focused on the specifics of the corresponding incident (Hassanzadeh et al, 2020) (Hemsley and Fisher, 2018) (Lee, Asante, and Conway, 2016). For instance, (Lee, Asante, and Conway, 2016) analysed the cyber-attack on the Ukrainian power grid. They

analysed technical components used by the attacker, attack mapped on to the ICS cyber kill chain to understand the attack elements, and opportunities to disrupt/restore. The information gathered from this review is used as the basis for developing the security incident management ontology model.

4. Effective Security Incident Management Ontology

In this section, the design requirements and choices considered to build the proposed model are described.

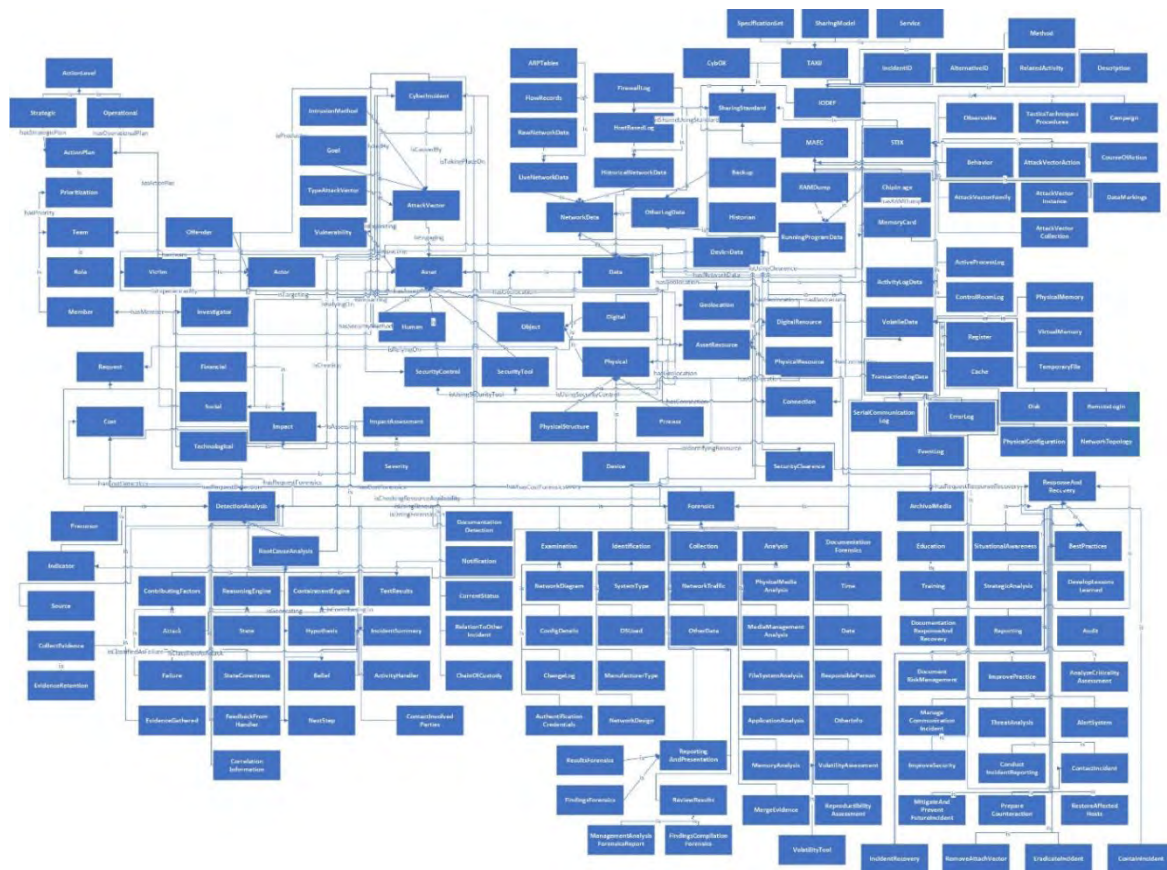


Figure 1: Model Architecture

This research considers the following design requirements (Dipert, 2013):

- to be humanly understandable by using controlled vocabularies.
- to be an ontology that uses widely known and accepted concepts.
- to be represented in one of the best available languages for formalizing ontologies, such as OWL or Common Logic; and
- to be able to apply methodologies for building ontologies and for illustrating instance-level data.

Further, as described in Section 2, the goal of the model is specified, the necessary information is collected and after that conceptualized and formalized in the fourth phase. Consequently, the architecture of the model containing the classes is illustrated in Figure 1, the global view of the model containing is depicted in Figure 2-left, the metrics of the model are captured in Figure 2-right, and the formal definition is provided in Equation (1) below:



Figure 2: Global view model (left) and its metrics (right)

Model = {UC, SC, I, AR}

(1)

In (1), UC represents the set of main classes of the model (i.e., first layer); SC represents the set of sub-classes of the model (i.e., second layer); I represent the set of instances / data values of the model (i.e., third layer); AR represents the set of attributes of classes and the relations between their instances.

From Figure 2-left, a series of different units with distinct colors are contained. The blue circles represent classes and sub-classes of the model, the blue rectangles represent the relations between the instances of these classes, the green rectangles represent the attributes of the classes while the orange rectangles represent the type of these attributes e.g., string or double. Furthermore, the followed layered perspective is considered to capture the elements of the model. On the first level are the upper classes (i.e., set UC) which extend the mother class Thing are located. These classes are Actor, AttackVector, Asset, ActionLevel, ActionPlan, Request, Team, SharingStandard, CyberIncident, Impact, Cost, DetectionAnalysis, Forensics, and ResponseAndRecovery. All other sub-classes (child classes) extend the upper classes.

On the second level, the child classes are located i.e., the sub-classes (i.e., set SC) that elaborate the concepts from the first level, such as Investigator (sub-class of Actor), DigitalResource (sub-class of Asset), IntrusionMethod (sub-class of AttackVector), ImpactAssessment (sub-class of DetectionAnalysis), Examination (sub-class of Forensics), ContainIncident (sub-class of ResponseAndRecovery), and Technological (sub-class of Impact). In Figure 3 are depicted a series of sub-classes of class ResponseAndRecovery which shows the processes and activities involved in the response and recovery phases of a cyber security incident.



Figure 3: Selection of Sub-classes of Class ResponseAndRecovery

On the third level, the instances (i.e., set I) that classes and sub-classes have, are represented. For exemplification purposes, in Figure 4, data of class Actor -> Investigator i.e. actors that have directly investigated this incident can be seen, response and recovery actions contained by class ResponseAndRecovery -> RestoreAffectedHosts, social effects that the incident had in class Impact -> Social, technological effects in class Impact -> Technological, and three types of requests for detection, forensics, and response and recovery actions that investigators put in place for the incident are contained in class Request.

On the fourth level, the attributes of classes are placed i.e., data properties (i.e., set AR) together with the relationships between them i.e., object properties. Accordingly, a selection them is further explained and depicted in Figure 5:

- *AnalysisStatus*: attribute of class *Analysis* that depicts the status of ongoing forensics analysis process and has two possible values: {on, off}.
- *EventLogPath*: attribute of class *EventLog* that contains the system path of the event log investigated and which is of type *string*.
- *IncidentSummaryAvailability*: attribute of class *IncidentSummary* which shows if the summary of an analysis conducted during the detection phase is available, and has two possible values: {yes, no}.
- *ResponseAndRecoveryDescription*: attribute of class *ResponseAndRecovery* that contains a brief description of the response and recovery phases of a cyber security incident, and which is of type *string*.
- *isAnalyzedBy*: relationship between class *CyberIncident* and class *Investigator* that shows the fact that a cyber security incident is analyzed by a security investigator e.g., defender who could be represented by an individual or a group of individuals in a security department or organization.
- *hasDeviceData*: relationship between classes *Physical*, *Digital*, *PhysicalResource*, *DigitalResource* and class *DeviceData* which shows which resource contain data retrieved from investigated devices.
- *hasRequestForensics*: relationship between class *Forensics* and class *Request* which shows the request done for starting the forensics process on a cyber security incident.

- *isSharedUsingStandard*: relationship between class *Data* and class *SharingStandard* to show how data is exchanged between two or more investigating entities.

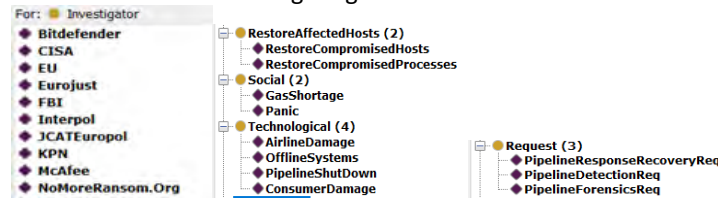


Figure 4: Instances or Data for the Colonial Pipeline Cyber Security Incident

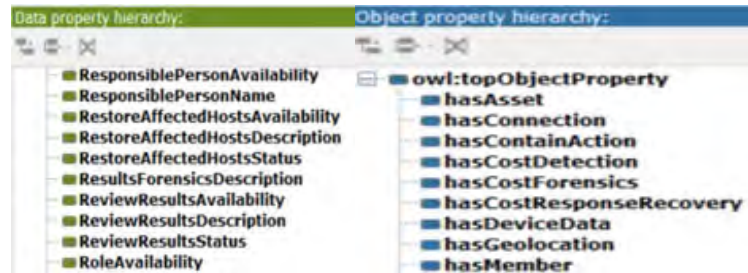


Figure 5: Selection of Attributes and Relationships in the Model

Determination of Layers in this Model: the first layer capture main phases, aspects, and agents involved in different processes of a cyber security incident; the second layer represent entities contained in the first layer with a focus on important activities and resources used; the third layer contain data-values or instances of the classes belonging to the first and second layer, and represent real or mimicked cyber incident data; the fourth layer embed characteristics of classes contained in the first and second layers that show attributes of these entities, and further captures relations between the instances or data-values corresponding to a cyber incident.

Until now, the results of the first seven phases of the methodological approach considered were presented i.e., the proposed implemented model. Furthermore, in the eighth phase the evaluation process takes place. Hence, the evaluation is carried out in two phases. First, by considering evaluation criteria such as consistency and reusability (Sawsaa & Lu, J 2012; Esposito, Zappatore & Tarricone, 2011). And second, through exemplification on a real-life ICS cyber incident. The evaluation from the first case is done in the development environment using the Hermit reasoner and was successful. The evaluation from the second case i.e., through exemplification is further presented.

Case Study Description: Colonial Pipeline operates over 5500 miles of pipeline extending from Texas to New Jersey in the United States (US) (TechTarget, 2021). It transports refined oil for gasoline, jet fuel, and home heating oil. It also directly serves jet fuel to about seven airports (TechTarget, 2021). Moreover, nearly 45% of east coast of US get their fuel from Colonial Pipeline. In May 2021, hacking group known as “DarkSide” carried out a ransomware attack on Colonial Pipeline (TechTarget, 2021). Ransomware is a type of malware which encrypts a victim’s files. The adversary might then demand ransom to restore access to the encrypted files upon payment. Even though this ransomware attack mainly affected the IT environment of Colonial Pipeline including its billing and accounting systems, this also disrupted fuel supplies as the operations were shut down as an immediate response to this ransomware attack and prevent the spreading of ransomware (TechTarget, 2021). This shut down led to panic buying, spike in gasoline price, and localized fuel shortage. This ransomware attack seems to be initiated through the stolen password for VPN (Reuters, 2021). Additionally, this legacy VPN system did not have multi-factor authentication, which made it easier for the adversary (Reuters, 2021). Personnel in the company seems to have used the same password for different accounts. One of which seems to be hacked earlier and the corresponding password was leaked in the dark web, which is then used for this purpose of getting into the VPN and carrying out the ransomware attack. DarkSide accessed the colonial pipeline network, stole 100 gigabytes of data within a two-hour window. Later, an employee noticed the ransom note and notified the operations supervisor. This led to the immediate shut down of operations. A security firm named “Mandiant” was involved in the investigation of this attack (TechTarget, 2021). Law enforcement and Government was also notified about this attack. The company paid the ransom to get access and reduce the recovery time (Reuters, 2021). In addition, at the later stage of the investigation, some of the ransom paid was recovered from the adversary group (Reuters, 2021). This ransomware attack was initiated on 7th May 2021 and the pipeline operations were restored on 12th May 2021 (TechTarget, 2021).

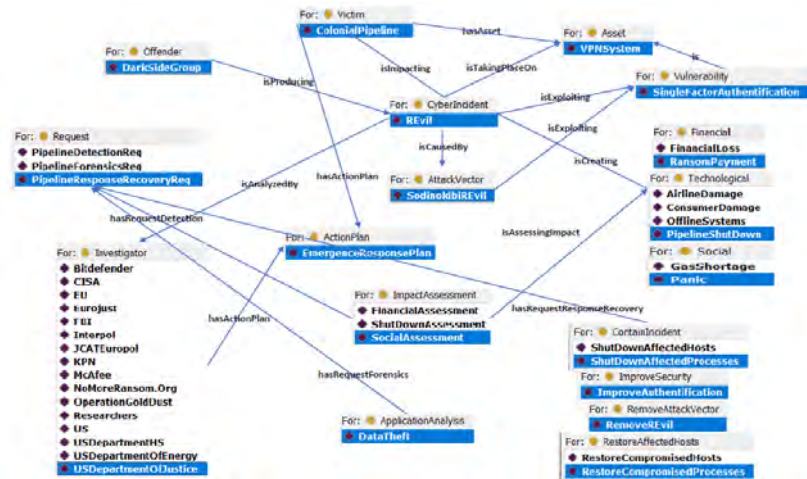


Figure 6: Colonial Pipeline Cyber Security Incident Data Visualization through Model Instantiation

To demonstrate the proposed model, the data obtained from the case study was populated in the model as depicted in Figure 6. This is compliant with the fourth design requirement (Dipert, 2013) considered when building the model, and further facilitates the following activities: (i) can serve as a common understanding basis for relevant stakeholders, (ii) can help to visualize relevant details of cyber security incidents during the mid-incident and post-incident phase for appropriately responding to an incident/learn relevant lessons from the incidents

5. Related Work

(Agrafiotis et al, 2018) advances a taxonomy of harm types that could be experienced by organizations from cyber-attacks. Furthermore, they have identified the stakeholders involved, their perceptions and priorities when dealing with the effects produced by cyber security incidents. (Arbanas & Čubrilo, 2015) provides a review of security ontologies based on a decade of scientific research until 2014. (Oltamari et al., 2015) developed an ontology for assessing human factors in cyber security by capturing individual characteristics, situational characteristics, and relationships influencing trust aiming at providing a useful basis for predicting and quantifying risk in required cyber risks assessments procedures. (Mundie et al., 2014) propose a meta-level ontological model for cyber security incident management which includes activities such as collecting evidence about the incident, analyzing the causes of the incidents, as well as drawing, and forming lessons learned.

(Mundie & Ruefle, 2012) discusses the current body of knowledge for cyber security incident management in a systematic approach. (Paul & Whitley, 2013) present a taxonomy for cyber security awareness questions that could support different activities taken while designing and developing cyber security awareness technologies. Among the aspects identified, the one named Damage Assessment is useful in this research as it refers to the response taken to effects occurred after these have been fully understood. (Onwubiko, 2018) advances an ontology for cyber security incident analysis. Notably, relevant to this research are the aspects involved in the need of escalation when managing cyber security incidents. (Maathuis et al, 2018a) propose an ontology that captures the essential entities and their corresponding relations in military Cyber Operations e.g., Effect which represents the impact that Cyber Operations have on targets and collateral entities. (Maathuis et al, 2018a) introduce a model for analyzing the effects of military Cyber Operations such as collateral damage and military advantage as a basis for military targeting decisions taken in Cyber Warfare. (Tagarev & Ratchev, 2020) propose a taxonomy for crisis management functions which although are not dedicated to the cyber security domain, they are still relevant to this paper. The authors identified a set of activities like making resources available to tackle an incident focusing on both short- and long-term response and recovery activities. Accordingly, both perspectives are integrated in the model proposed.

6. Conclusions and Future Work Directions

Efficacious security incident management plays a crucial role in lessening the negative consequences of an incident in terms of organization's economy, reputation, and personnel safety in addition to recovering/restoring back to normal operations after the incident in a short down time. There are various concepts that are proposed in addition to globally accepted standards like CyBOX, STIX, TAXII in the scientific literature relevant for security incident management. However, a holistic overview of such concepts and their

relationships that could help to operationalize them during incident response is missing. Therefore, in this paper, we conducted an extensive literature review on security incident management especially covering the mid-incident and post-incident phases. This is categorized as four different themes including root cause analysis. Information sharing, forensics, and lessons learned. Based on the extensive literature review, a security incident management ontology model is developed with relevant classes and sub-classes. Firstly, the developed model is evaluated using Hermit reasoner. Additionally, the model is populated with the relevant information from the colonial pipeline cyber security incident to show how this developed model can be used during the post-incident phase of security incident management. In the future, using the proposed approach, the model can be further extended with additional information from the relevant grey literature like white papers, standards, technical reports which might be more applicable for security incident management. This model represents the basis for the development of Artificial Intelligence (AI)/Machine Learning (ML)-based models. For instance, in root cause analysis, the developed ontology model can provide inputs in terms of variables and their relationships to the structural development of AI/ML-based models that would help identify the root cause and select effective response strategies. Moreover, the effectiveness of the developed security incident management ontology model needs to be evaluated in the future during an ongoing incident/mock incident.

References

- Alrimawi, F., Pasquale, L., Mehta, D. and Nuseibeh, B., 2018, May. I've seen this before: Sharing cyber-physical incident knowledge. In *Proceedings of the 1st International Workshop on Security Awareness from Design to Deployment* (pp. 33-40).
- Cichonski, P., Millar, T., Grance, T. and Scarfone, K., 2012. Computer security incident handling guide. NIST Special Publication.
- Chockalingam, S., Pieters, W., Teixeira, A., Khakzad, N. and Gelder, P.V., 2018. Combining Bayesian networks and fishbone diagrams to distinguish between intentional attacks and accidental technical failures. In *International Workshop on Graphical Models for Security* (pp. 31-50).
- Chockalingam, S. and Katta, V., 2019, December. Developing a Bayesian Network Framework for Root Cause Analysis of Observable Problems in Cyber-Physical Systems. In *2019 IEEE Conference on Information and Communication Technology*.
- Chockalingam, S., Pieters, W., Teixeira, A., and van Gelder, P., 2021. Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates. *Cybersecurity*, 4(1), pp.1-19.
- CIDAWG, 2015. Enhancing Resilience Through Cyber Incident Data Sharing and Analysis.
- Cooper, C., Booth, A., Varley-Campbell, J., Britten, N. and Garside, R., 2018. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC medical research methodology*, 18(1), pp.1-14.
- Esposito, A, Zappatore, M & Tarricone, L 2011, 'Evaluating scientific domain ontologies for the electromagnetic knowledge domain: A general methodology', *Journal of Web & Semantic Technology*. vol. 2, no. 2, pp. 1-19.
- Fernández-López, M, Gómez-Pérez, A & Juristo, N (1997) 'Methontology: From ontological art towards ontological engineering', *Proceedings of the fourteenth national conference on Artificial Intelligence, AAAI-97*, Spring Symposium Series, pp. 33-40.
- Findrik, M., et al., 2017. Trustworthy Computer Security Incident Response for Nuclear Facilities.
- Goodwin, C., Nicholas, J.P., Bryant, J., Ciglic, K., Kleiner, A., Kutterer, C., Massagli, A., McKay, A., Mckitrick, P., Neutze, J. and Storch, T., 2015. A framework for cybersecurity information sharing and risk reduction. *Microsoft*.
- Hassanzadeh, A., Rasekh, A., Galelli, S., Aghashahi, M., Taormina, R., Ostfeld, A. and Banks, M.K., 2020. A review of cybersecurity incidents in the water sector. *Journal of Environmental Engineering*, 146(5), p.03120003.
- He, Y., Inglut, E. and Luo, C., 2022. Malware incident response (IR) informed by cyber threat intelligence (CTI). *Science China Information Sciences*, 65(7), pp.1-3.
- Hemsley, K.E. and Fisher, E., 2018. *History of industrial control system cyber incidents* (No. INL/CON-18-44411-Rev002). Idaho National Lab, United States.
- Hevner, AR, March, ST & Park, J. (2004), 'Design research in information systems research', *MIS Quarterly*, vol. 28, no. 1.
- Kulikova, O., Heil, R., van den Berg, J. and Pieters, W., 2012, December. Cyber Crisis Management: A decision-support framework for disclosing security incident information. In 2012 International conference on cyber security (pp. 103-112). IEEE.
- Lallie, H.S., et al., 2021. Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *Computers & Security*, 105, p.102248.
- Lee, R.M., Assante, M.J., Conway, T., 2016. Analysis of the cyber-attack on the Ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388.
- Lee, J.W., Song, J.G., Son, J.Y. and Choi, J.G., 2018. Propositions for Effective Cyber Incident Handling.
- Luijff, et al. "On the sharing of cyber security information." *International Conference on Critical Infrastructure Protection*. 2015.
- Maathuis, C., Pieters, W. & van den Berg, J. (2018a). 'A Computational Ontology for Cyber Operations'. In *Proceedings of the 17th International Conference on Cyber Warfare and Security, ICCWS*, pp. 278-88.

- Maathuis, C., Pieters, W., & van den Berg, J. (2018b). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.
- Mavroeidis, V., & Bromander, S. (2017). Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pp. 91-98.
- Mundie, D. A., & Ruefle, R. (2012, August). Building an incident management body of knowledge. In *2012 Seventh International Conference on Availability, Reliability and Security*. IEEE (pp. 507-513).
- Mundie, D. A., et al. (2014, January). An Incident Management Ontology. In *STIDS* (pp. 62-71).
- Okoli, C., 2015. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37(1), p.43.
- Oltramari, A., Henshel, D. S., Cains, M., & Hoffman, B. (2015). Towards a Human Factors Ontology for Cyber Security. *Stids*, 2015.
- Onwubiko, C. (2018). Cocoa: An ontology for cybersecurity operations centre analysis process. In *2018 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA)* IEEE (pp. 1-8).
- Paul, C. L., & Whitley, K. (2013). A taxonomy of cyber awareness questions for the user-centered design of cyber situation awareness. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer (pp. 145-154).
- Peffer, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.
- Piatkowska, E., et al., 2020. Online Reasoning about the Root Causes of Software Rollout Failures in the Smart Grid. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* (pp. 1-7).
- Pranggono, B. and Arabo, A., 2021. COVID-19 pandemic cybersecurity issues. *Internet Technology Letters*, 4(2).
- Rajamäki, J., Tikanmäki, I. and Räsänen, J., 2019. CISE as a Tool for Sharing Sensitive Cyber Information in Maritime Domain.
- Reuters, 2021, One password allowed hackers to disrupt Colonial Pipeline. <https://www.reuters.com/business/colonial-pipeline-ceo-tells-senate-cyber-defenses-were-compromised-ahead-hack-2021-06-08/>
- Sawsaa, AF & Lu, J 2012, 'Building information science ontology (OIS) with Methontology and Protégé', *Journal of Internet Technology and Secured Transactions*, no. 1, vol. 3/4.
- Shinde, N. and Kulkarni, P., 2021. Cyber incident response and planning: a flexible approach. *Computer Fraud & Security*, 2021(1).
- Spyridopoulos, T., Tryfonas, T. and May, J., 2013. Incident analysis & digital forensics in SCADA and industrial control systems.
- Staves, A., Balderstone, H., Green, B., Gouglidis, A. and Hutchison, D., 2020, May. A Framework to Support ICS Cyber Incident Response and Recovery. In *the 17th International Conference on Information Systems for Crisis Response and Management*.
- Stirland, J., et al., 2014, October. Developing cyber forensics for SCADA industrial control systems. In *The International Conference on Information Security and Cyber Forensics (InfoSec2014)*. pp. 98-111.
- TechTarget, 2021. Colonial Pipeline hack explained: Everything you need to know. <https://whatis.techtarget.com/feature/Colonial-Pipeline-hack-explained-Everything-you-need-to-know>
- Van Vliet, P., Kechadi, M.T. and Le-Khac, N.A., 2015. Forensics in industrial control system: a case study. In *Security of Industrial Control Systems and Cyber Physical Systems* (pp. 147-156).

Bureau of Justice Assistance Student Computer and Digital Forensics Education Opportunities Program: The Assessment of Online Graduate Students

Kyung-Shick Choi¹, Chitkushev Lou², Kyung-Seok Choo³ and Claire Seungeun Lee⁴

¹Professor of Practice, Department of Applied Social Science, Metropolitan College, Boston University, USA

²Associate Professor, Computer Science, Associate Dean, Metropolitan College, Boston University, USA

³Professor, School of Business & Justice, Utica College, USA

⁴Assistant Professor, School of Criminology & Justice Studies, University of Massachusetts Lowell, USA

kuung@bu.edu

lrc@bu.edu

kychoo@utica.edu

claire_lee@uml.edu

Abstract: The current capabilities of many law enforcement agencies are tightly constrained despite the heightened level of awareness and concern for the role recent technology has in facilitating cybercrime and instances of online victimization. More specialized computer forensics and digital evidence training programs are necessary to meet the needs of local and state law enforcement agencies. Based on the context, this paper discusses an interdisciplinary approach to addressing this dilemma while providing in-depth computer forensics and cybercrime investigation training that is both informative and pragmatic to future law enforcement officers. Using pre- and post-test results, this study assesses students' technical background levels, reflecting comprehensive course learning objectives and pre-training levels of applied digital forensic investigation knowledge. Results suggest that students' technical abilities and knowledge of different investigative tools significantly improved after the program. In particular, the program not only strengthened students' knowledge of digital forensic investigation, but also helped students achieve higher t-test scores. We expect our study results to provide recommendations for cyber programs in other higher education institutions. The findings will serve as a guide for enhancing the current capacities of other higher education institutions to better serve their students in areas of computer forensics and digital evidence. In the long term, these efforts will lead to more effective cybercrime investigation and successful prosecutions, ultimately reducing cybercrime victimization.

Keywords: computer forensics, training, law enforcement, cybercrime investigation

1. Introduction

As demonstrated by a string of recent cases, cybercrime problems have proliferated as digital and information technology become more widely adopted. Throughout the COVID-19 pandemic, cybercriminals have utilized mass fear and panic to exploit vulnerable victims through ransomware attacks, counterfeit remedies, coronavirus phishing, federal stimulus check scams, and fake charities for coronavirus causes.

In the digital age, awareness of and concern about the role of the latest technology in facilitating cybercrime and online victimization has grown tremendously. Unfortunately, the current capabilities of many law enforcement agencies to investigate such crimes remain limited, as most officers lack the necessary training in processing computer data and related evidence to effectively investigate cybercrime. To meet the rapidly digitizing demands of state and local law enforcement, more specialized computer forensics and digital evidence training programs are required.

The current study explores in-depth computer forensics and cybercrime training designed to be both informative and pragmatic about the students currently enrolled in Criminal Justice (CJ) and cybercrime/computer science (CS) program in a higher education institution and who might be future law enforcement officials. We use an interdisciplinary approach to assess the effectiveness of this program. This study uses a mixed-methods research design and evaluates student technical training, reflecting the comprehensive learning objectives of the course and the level of knowledge in the applied digital forensic exams prior to the online training. The results of the pre- and post-tests demonstrate that students' technical skills and knowledge of various digital forensics tools

improved significantly after learning the tools in the program. The study was especially effective at strengthening students' knowledge of digital forensic examinations and helping students achieve higher test scores in our post-test.

2. Related work

Originating in the late 1980s, law enforcement professionals used the term computer forensics to refer to the examination of independent computers for digital criminal evidence (Yasisac et al, 2003). Computer/digital forensics later evolved to describe the post-incident analysis of computer attacks by intruders or malicious code (Liu and Wang 2009). The emergence of cybercrime and cybersecurity issues gave rise to the expectation that law enforcement agencies (LEA) should equip themselves with computer forensics knowledge and practical skills. However, there are currently at least two issues with the LEAs' approaches to computer forensics. On the one hand, LEAs have vastly differing levels of resources, capacity, and manpower. Small local LEAs are not as advanced in these regards. On the other hand, despite that computer forensics is an emerging field, much complicated and higher-level training in computer forensics is also in development. While most of the programs are focused on computer science, a limited number of programs are offered for other relevant social sciences (e.g., criminology/criminal justice).

Digital forensics training at higher education institutions offers students the opportunity to acquire skills, exposes them to hands-on exercise, and allows them to gain new knowledge. In doing so, curriculum development to accommodate recent skills and experiences for students is important (Antunes and Rabadao 2019). Most of the existing studies remain focused on undergraduate computer science/cybersecurity degrees (Chi et al, 2009; Ward 2020); however, these studies give us universal insights that can be applied to a postgraduate level. Zhang and Choo (2020) suggest an experiential learning approach for successful digital forensic education. Hands-on labs and case studies are an effective pedagogical method of teaching cybersecurity (Du and Wang 2009; Flores et al, 2021; Yuan et al, 2014; Ward 2020). A scenario-based and "learning by doing" approach is effective (Batten and Pan 2008; Ward 2020). Experience in real-world problems plays an important role of students' pathway to employment (Mison et al 2020).

In an examination of online education in computer and digital forensics, Kessler (2007) found that online courses with active, problem-based learning methods are better suited to the needs of adult learners (Kessler 2007b). Training, teaching students, and nurturing their technology knowledge base is critical. Most of all, the competence of fundamental computer skills and the capacity to demonstrate practical skills are essential in computer/digital forensics (Anderson et al., 2006; Antunes and Rabadao 2019). In particular, online labs provide one of the most supportive environments for active learning, as technology-transparent courses are more interactive, translating to successful experiences for learners (Kessler 2007b; Flores et al, 2021). Online formats create an opportunity to students to catch up material (Buchanan et al, 2010).

3. Current study, data, and methods

3.1 Curriculum design

To provide students with the essential computer forensics and digital evidence skills that will prepare them to work within all levels of law enforcement agencies, the research team conducted expert focus groups and steering committee meetings with representatives from local, state, and federal law enforcement agencies. The experts reviewed the contents of relevant courses, including techniques and software, learning objectives, and assessment tools to determine whether they meet the current needs of law enforcement. Most class projects of the finally selected courses for the study consist of practical labs and hands-on exercises that strongly tie into the areas of computer forensics and cybercrime investigation, and are aligned with the best practices based on current law enforcement training manuals, CDFE (Certified Digital Forensic Examiners) curriculum mapping, and NIST (National Institute of Standards and Technology) Digital Forensic Techniques Guide.

Due to the COVID-19 pandemic, all existing courses were converted to the online format. We assessed the program goals based on the program curriculum map below. It involved (1) course objectives' alignment with the program's goals, (2) a systematic review of course syllabi, textbooks, and teaching materials, and (3) course-embedded assessment using oral presentation, paper, projects, lab practice, and graded thread discussions. We also developed pre/posttest course assessment tools and implemented to achieve a practical course level of assessment.

We used the Cybercrime (CJ 610) course to assess the introduction levels of the program goals, while the Applied Digital Forensic Investigation (CJ710) course determined the reinforcement levels. For the proficiency level, we used the Digital Forensics and Investigations (CS693) and Mobile Forensics and Security (CS694)¹ courses to determine the demonstrated achievement of program goals. We used the curriculum map as a guide to assess whether each class introduced, reinforced, and provided the appropriate levels of the program goals. We utilized various assessment tools such as TD (thread discussions), PP (papers), PJ (project), C (capstone), and LB (lab).

The intent of our assessment was to develop a refined cybercrime investigation training curriculum for future law enforcement officers. This assessment provided students with the most comprehensive and reliable cybercrime investigation training available in regard to Computer Forensics and Digital Evidence Educational Training.

	1. Demonstrate an understanding of key concepts and processes in computer forensics and digital evidence.	2. Demonstrate critical thinking and analytical performance capability related to computer forensics and digital evidence.	3. Use various tools to conduct computer forensics and analyze digital evidence.	4. Apply appropriate legal and ethical concepts to computer forensics and digital evidence.
Courses				
CJ 610 Cybercrime	I	I	I	I
CJ 710 Applied Digital Forensic Investigation	R	R	R	R
CS 693 Digital Forensic Investigation	P	P	P	P
CS 694 Mobile Forensics and Security	P	P	P	P

Abbreviations. **I = introduced. R = reinforced. P = proficiency.**

I = Introduced; student learning outcomes will be measured if the course introduced relevant program goals

R = Reinforced; student learning outcomes will be measured if the course provided practice opportunities

P = Proficient; student learning outcomes will be measured for the demonstrated achievement of the program goals.

Figure 1: Curriculum design

3.2 Mixed method research design and data collection

We used a mixed-method research design, which originated in the social sciences and is used in other disciplines. This approach is based on comprehensive data collection and provides similarly robust and multi-faceted results. This study integrates quantitative and qualitative research and follows an explanatory sequential design, also known as an “explanative design.” It consists of two phases: quantitative data collection and analysis, followed by qualitative data collection. The latter is used to explain the initial quantitative results, also known as a “Quat + Qual” model (Schoonenboom and Johnson 2017). We essentially use the same sample to extract two types of data in succession.

3.3 Quantitative method

All students who contributed data for this study participated in pre-post online surveys to assess the Student Computer Forensics and Digital Evidence Educational Opportunities Program. The quantitative analysis aims to (1) assess the level of students’ technical background, per the course’s comprehensive learning objectives and level of knowledge in the applied digital forensic investigation prior to the online training, and (2) evaluate the impact of the training by comparing pre-and post-test results.

¹ The course numbers and titles indicate: Criminal Justice 610: Basic Technical Level, Criminal Justice 710: Basic-Intermediate Technical Level, Computer Science 693 Intermediate-Advanced Technical Level, and Computer Science 694: and Advanced Technical Level.

Questionnaires were designed by two researchers and cross-validated by a third person. Participants are students in a Master's in Criminal Justice program, and online surveys were sent to participants during the first day of class and the last week of each semester. For the quantitative pre- and post-test data collection segment, we analyzed scores from four core courses offered in Spring 2020 and Summer 2021. Qualified students were selected as a group, and the participants were responsible for completing pre-post testing and contributing interviews for program evaluation purposes. To conduct an overall assessment for the Student Computer Forensics and Digital Evidence Educational Opportunities Program, all students were encouraged to participate in the four pre-post online surveys.

We performed a paired t-test to determine the difference between dependent variables for the same subject in two related groups. We compare pre-class and post-class test scores to examine the effectiveness of the four courses taken by participants during their program in increasing technical background and knowledge related to digital forensics. Each questionnaire had two parts: the "student technical background measure" and the "knowledge test." Concerning CJ710 and CS693, participants filled out a 16-item questionnaire (student technical background measure) and a 20-item questionnaire (knowledge test) before and after the training. For the third course (CS694), participants filled out a 13-item questionnaire (student technical background measure) and a 28-item questionnaire (knowledge test) before and after the training. For CJ610, participants filled out a 15-item questionnaire (student technical background measure) and a 20-item questionnaire (knowledge test) before and after the training. Each pre- and post-test asked the same set of questions, varying only by course, and students were asked to state their confidence levels regarding meeting course learning objectives. Only participants who completed both pre- and post-tests were included in the analysis (CJ710: N=31, CS693: N=32, CS694: N=20, CJ610: N=23). Using the pair-sample t-test, the differences between pre- and post-test scores were examined for statistical significance.

3.4 Qualitative method

After completing the pre-post tests, four students in each class were selected to partake in qualitative data collection. The qualitative analysis consisted of conducting interviews to assess the effectiveness of the program's four courses: (1) CJ610, (2) CJ710, (3) CS693, and (4) CS694. Specifically, the interview questions were developed in effort to reflect course objectives, and the responses were utilized to assess the courses' areas of success and those where improvements are needed. These participants were randomly selected using the program SPSS seed function. Each student had an equal chance of being assigned to an interview group, and each interview was held online via Zoom for approximately one hour. We recorded interviews of participants who consented and otherwise jotted down key information and details within a day of interview completion.

4. Results

4.1 Quantitative analysis

We used a series of paired t-tests to evaluate the overall effectiveness of each course by determining a statistical difference between the means of the pre- and post-test scores. For the first question/hypothesis on student technical background differences, the finding suggests that the average pre-test score for student technical background (CJ710: M=32.10, SD=13.66; CS693: M=31.69, SD=11.11; CS694: M=22, SD=9.64; CJ610: M=36.27, SD=13.64) is significantly lower than the post-test score (CJ710: M=67.06, SD=9.84; CS693: M=62.59, SD=11.08; CS694: M=57, SD=7.66; CJ610: M=61.91, SD=9.60). Here, the differences between the pre-post test scores indicate that the training contributes to the level of student technical background and knowledge about digital forensic investigation. As expected, student technical knowledge mean scores are more likely to increase after completing each course, as students gain more knowledge over the semester.

Likewise, the findings for the second measure suggest that the average pre-test score for knowledge in digital forensic investigations (CJ710: M=5.52, SD=3.40; CS693: M=8.28, SD=4.10; CS694: M=6.53, SD=5.58; CJ610: M=6.13, SD=3.35) is significantly less than the post-test score (CJ710: M=15.58, SD=2.88; CS693: M=14.13, SD=3.61; CS694: M=21.67, SD=3.22; CJ610: M=12.74, SD=2.50). In other words, due to the training that students received over the duration of the semester, mean test scores measuring knowledge about the digital forensic investigation are more likely to increase after course completion.

Overall, these results provide quantitative evidence that training (via course completion) influences the difference between pre- and post-test scores. More specifically, mean test score results from the pre- and post-

tests demonstrate that levels of student technical background and levels of digital forensic investigation knowledge are strengthened after course completion.

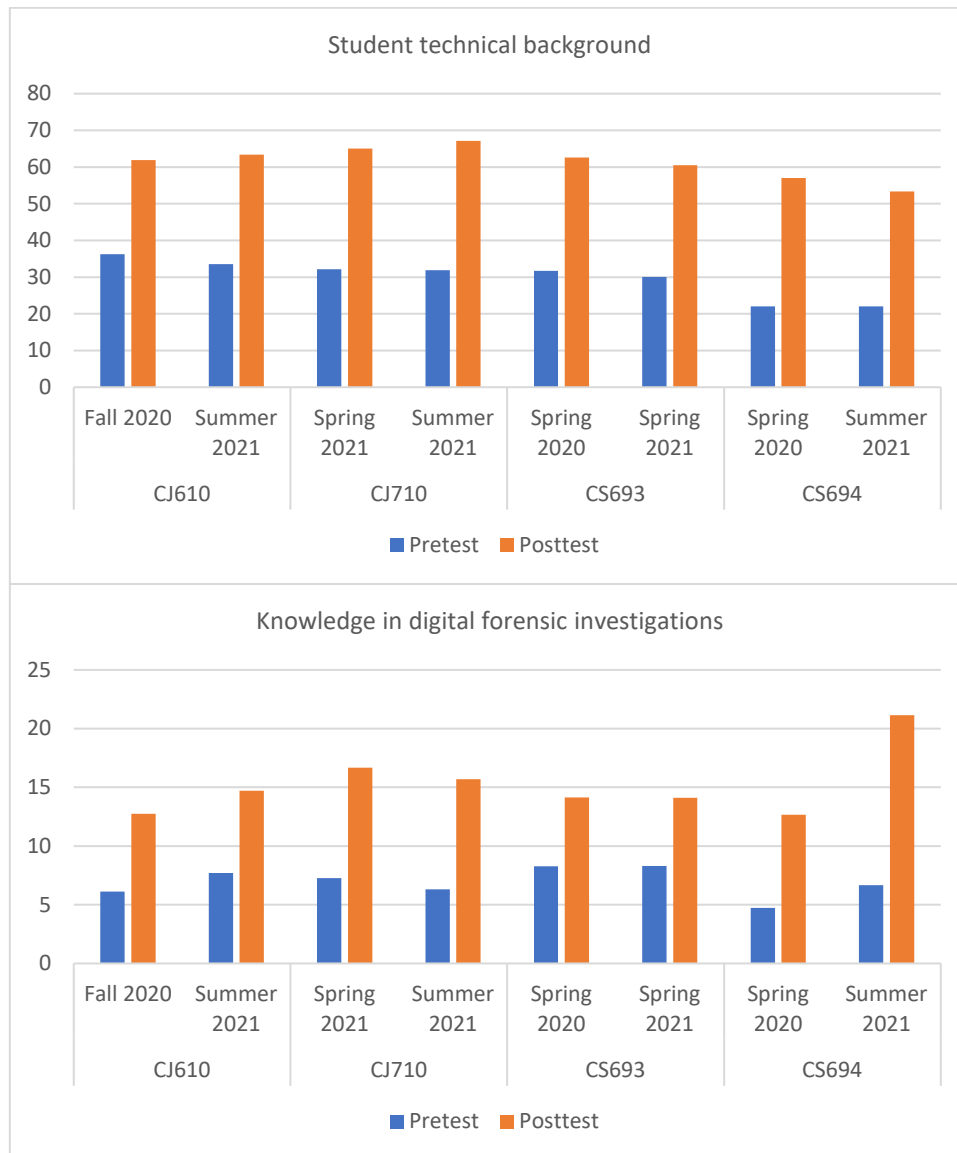


Figure 2: Pre-post result

Note: Scores indicate means.

4.2 Qualitative analysis

Computer forensics is one of the core themes of cybercrime investigation. However, this area has yet to be introduced widely in the CJ program, even though it has been taught extensively in the CS program. Given that our participants are primarily enrolled in the CJ program and work in former, current, or future law enforcement sectors, this training in computer forensics plays an important role in shaping their learning experience, confidence, and career. In what follows, we abide by the general structure of program evaluations and analyze interviews in relation to the quality of training course instruction, training content, and online training.

4.3 Quality of training course instruction

Participant 3 (CJ710) explained that,

A theme throughout the course was cybercrime investigations utilizing a forensic tool known as, EnCase. Students were provided with an interesting storyline following a suspect named Mr. Evil, providing his students with just enough information, so even the experienced people are given the opportunity to dig in deeper and explore the material like a treasure hunt.

Even though some of the interviewees already had relatively strong backgrounds and positive experiences working in the field of computer science, they enjoyed how the course instructors were consistently available to provide guidance, explain everything step-by-step, and take students back to the basics in efforts to teach the fundamental purpose behind everything that was being taught in the course.

Participant 1 (CJ710) appreciated how the course calendar was readily available and shared in advance so that students were always aware of course due dates, group discussions, virtual labs, and everything else that was essential to succeeding. He further explained that even when something came up in their personal life, “The instructor was very accommodating and understanding.” As mentioned in the interview,

Even when I did not ace the assignment, I always received great feedback as to why I received that mark and how I could have completed it a little bit differently, which I think is the most important thing. Either way, you are able to know what you got right and what you needed to improve on.

Participant 4 (CS693) offered similar feedback, explaining that assignments were accompanied by sufficient instructions throughout the course. As stated in the interview:

[T]he professor made himself available with office hours and a teaching assistant as a technical resource who was available for any software issues that students were experiencing. The assistant assisted me in loading necessary software onto my laptop. The professor also posted lecture videos for weeks that we did not meet. I found the professor to be incredibly prepared, qualified, and fair in his grading practices.

Overall, participants explained that these courses piqued their interest. They consistently felt challenged. They were pleased with the amount of new material (e.g., digital forensic tools) they learned to use, which they could apply in their specific careers. In addition, participants indicated that they were excited to learn more and expressed that they feel measurably more confident as well as prepared to continue onto their future endeavors.

4.4 Quality of training content

The next set of interview questions asked participants to rate the quality of the course content on how well it provided them with a greater understanding of the material and increased their ability to think critically.

A majority of the participants’ responses described how consistent the quality of the content was and how the organization of the material was vital to helping them meet their learning outcomes. The content and different topics demonstrated weekly progression, and participants felt they were always provided the opportunity to go back and fully relearn challenging material. According to Participant 3 (CJ710):

[T]he instructor break (sic) up each module, so there was not consistent overlapping. Each module was well thought out and each was broken up according to the specific topic, making each module unique.

Furthermore, all participants appreciated that they were able to dive into the course material headfirst and hands-on, expressing how grateful they were to have the opportunity to work with tools and programs that law enforcement agencies utilize on a daily basis.

Participant 4 (CJ710) explained how essential application learning truly is, especially when it comes to certain technical skills. Participant 4 (CS693) provided similar feedback and further emphasized that this course gave him a greater understanding of how digital forensic investigations are conducted. More specifically, the participant indicated that, “from learning about hexadecimal mathematical equations to hardware architecture and design to the forensic processing of digital evidence files; this course has certainly enhanced my understanding of the core concepts” (Participant 4, CS693). As Ward (2020) found, our study results show working hands-on with the software utilized in class was found to be much more effective than traditional textbook learning.

As mentioned in their interview, Participant 2 (CJ710) explained how this course truly, “enhanced my critical thinking, especially within our discussions because each student needed to consider their statements and their responses to others. Hearing these different perspectives was eye-opening because when a topic was up for discussion, students often approached it in a way that others never even thought about.”

Relatedly, Participant 4 (CJ710) explained that in some of the lab exercises, students would have to figure out where certain evidence was located. This participant continued by explaining that, “I could apply that as a critical thinking exercise in order to decide if it truly mattered and if it was relevant to the case.” Participant 4 (CS694) provided similar feedback by stating that he appreciated how this course provided him, “with a greater understanding of mobile forensics, security, and the challenges examiners are often faced with when extracting data from these devices.” Similarly, Participant 2 (CJ610) explained that she definitely believes her

[K]nowledge/skills have improved by taking this course. I now find myself instinctively using the jargon that I have learned and begun educating and aiding those around me in becoming more mindful about vulnerabilities and exploitations.

As in any field of study, providing students with the opportunity to practice their learned skills and acquired knowledge has proven to be a key component in achieving notable accomplishments and future success. As one can see, the opportunity to work hands-on with real-life examples/cases can truly positively impact students’ learning outcomes.

4.5 Quality and advantages of the online training

When asked questions regarding the advantages of the online modules and lessons along with the courses’ overall learning outcomes, a majority of participant responses reflected upon how flexible the online course schedule was for them. All participants are graduate students and have full-time jobs with busy home life and other obligations that they must attend to on a daily basis. Thus, having the flexibility to complete assignments and review course content on their own time was extremely beneficial. More specifically, participants explained that they enjoyed the course content and how each module/lesson began with the basics, eventually progressing to more advanced topics involving in-depth discussions and opportunities to grasp new material fully. All participants much appreciated the convenience of easily accessing assignments and navigating the site.

Participant 3 (CJ710), who currently works as a computer forensic examiner, explained that the course outcomes exceeded his expectations. This participant stated in his interview that, “As my full-time job for over a year now, I did not think that I would learn too much more. However, because I never used EnCase before this course, I actually got some good use out of it, and I learned a lot. I did not expect to learn anything new, but I truly did.” Participant 4 (CS693), who has completed over twenty years of service as a police officer, explained that the course outcomes were consistent with his expectations. This participant explained that having done hundreds of criminal investigations of cybercrimes, he had some familiarity with the unique challenges that this field may bring. According to the participant:

[T]his course absolutely enhanced my knowledge of cybercrime. To effectively investigate cybercrimes, investigators must have a basic knowledge of digital devices and must work collaboratively with forensic examiners in these types of cases. I learned the language of digital forensics, the limitations, and the importance of proper scope of searches are all areas within this online course.

Participant 2 (CS693) shared a similar learning experience and is currently employed in cybersecurity, specifically performing cybercrime investigations. In the interview, the participant explained that she “enjoyed the self-paced learning of the online modules and the different types of interactions they provided (videos, reading, etc.). I achieved the overall learning goals, and they truly exceeded my expectations. The Digital Forensics course is one of my most favorite courses that I have taken during my graduate studies.” The benefits of online courses that were discussed by the student were in line with Buchnnan et al. (2020)’s study.

Relatedly, Participant 4 (CS694), who also has impressive credentials in the field, stated in the interview that, “This was a far more technical course in the core workings of cellular phones. However, I gained valuable knowledge by learning where to find items of evidentiary values on cell phones and the value of using a variety of tools to extract data from cell phones.”

As can be seen, even participants with years of background experience from working in cyber-related fields were still able to learn new information from this course and were fortunate to apply this acquired knowledge to their careers and future endeavors. Overall, participants felt that this course was well structured and highly organized, and they were happy to report that they achieved the overall learning goals and objectives.

5. Discussion and conclusion

This study aimed to understand and evaluate the effectiveness of computer forensics and digital evidence educational training with a mixed-methods design. The study assessed students' technical background levels, reflecting course comprehensive learning objectives and levels of knowledge surrounding applied digital forensic investigations prior to the online training. The study evaluated the impact of the training by looking at both the differences between pre- and post-test scores and interview data. Existing research suggests that pre-post learning measures are helpful to establish students' baseline abilities and evaluate the effectiveness of teaching methods (Flores et al, 2021; Stark-Wroblewski et al, 2007). Quantitative pre- and post-test results were integrated with and further explained by qualitative research findings. Using an explanatory sequential design, a combination of quantitative data collection and analysis followed by the collection of qualitative data, the second layer of qualitative data is used to explain the initial quantitative results.

Pre-post test results functioned as numeric representations of students' technical backgrounds, acquired knowledge, and practical skillsets. The study's results indicated that the students, who lacked relevant technical backgrounds, became confident about the subject matter and gained substantial technical knowledge. These findings provide evidence that a carefully designed curriculum for computer forensics and digital evidence educational training can improve students' knowledge, skillsets, and competency. These benefits can help adequately prepare students for real-world scenarios.

Based on a mixed-methods research design following the development of four courses over two years, this research contributes to how a cybercrime investigation curriculum can be refined for future law enforcement officers training. The project goal is to provide students with the most comprehensive and reliable cybercrime investigation training available in Computer Forensics and Digital Evidence Educational Training.

Furthermore, this research provides implications for future research and curriculum development. As technology rapidly evolves and is just as quickly adopted by cybercriminals, training in cybercrime investigation should also be preemptively pursued, increasing responsiveness to changes in the field. Training programs should reflect the benefits and challenges of evolving technology and include more hands-on exercise. More importantly, programs should develop up-to-date training modules with hands-on coursework using core techniques that mirror real-life scenarios. As this study affirmed, hands-on training is an effective tool for trainees (Beuran et al, 2018; Liu 2016; Wiedenbeck and Zila 1997). Scenario-based learning, onsite triage training, and firsthand exercises are some of the most important techniques involved in software training and cybersecurity training (Beuran et al. 2018; Gonzalez-Manzano and De Fuentes 2019). Such immersive experience enhances students' motivation (Olfman and Bostrom 1991), understanding of software, technical knowledge, skills, and perceived utility of what they are learning (Chisholm 2015; Kessler 2007a). More specifically, Rosenshine (2012) suggests that hands-on exercises are beneficial for students' information-processing, which significantly increases the chances of long-term memory retention for newly learned material.

In the long run, new curricula are necessary to address the rapidly evolving landscape of cybercrime and cybersecurity (Choi, 2015; Choi et al., 2020). AI would be one of the most important drivers and disrupters of changes in cybersecurity and cybercrime investigation. As demonstrated in Blackhat conference past years, AI can be used as an attack vector for hacking practices via AI-enhanced malware (Charan, 2019). Internet of Things (IoT) is being adopted in many areas of every-day life and becoming a hot topic around the world. With an expected growth of exponential measures, IoT brings the promise of generating huge revenues. In return, society faces significant challenges from rising cyber threats such as social engineering and phishing attacks via IoT based Botnets (Yaqoob et al., 2017). Curricula on the Internet of Things, Artificial Intelligence and other emergent technologies that cybercriminals are equipped with need to be further developed in the future.

References

- Abeldina, Z., Moldumarova, Z., Abeldina, R. and Moldumarova, Z.I. (2015). "Virtual Learning Space in the System of E-Learning", *Mediterranean Journal of Social Sciences*, Vol 6, No. 5, p. 478.
- Anderson, P., Dornseif, M., Freiling, F. C., Holz, T., Irons, A., Laing, C. and Mink, M., (2006). "A Comparative Study of Teaching Forensics at a University Degree Level", In: Göbel, O., Schadt, D., Frings, S., Hase, H., Günther, D. & Nedon, J. (Hrsg.), *IT-Incident Management & IT-Forensics - IMF 2006*. Bonn: Gesellschaft für Informatik e. V. (pp. 116-127).
- Antunes, M. and Rabadao, C. (2019). "Cybersecurity and Digital Forensics – Course Development in a Higher Education Institution", *Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPar 2018)*, pp. 338-348.

- Batten, L. and Pan, L. (2008). "Teaching digital forensics to undergraduate students", *IEEE Secur Priv*, Vol 6, No. 3, pp. 54-56.
- Buchanan, B., MacFarlane, R., and Ludwiniak, R. (2010). "Student perception of on-line lectures within a blended learning environment for security and digital forensics".
- Beuran, R., Tang, D., Pham, C., Chinen, K.-I., Tan, Y. and Shinoda, Y. (2018). "Integrated framework for hands-on cybersecurity training: CyTrONE", *Computers & Security*, Vol 78, pp. 43-59.
- Charan, H. (2019, April 22). Use of ML and AI in Blackhat hacking. Medium. Retrieved October 26, 2021, from <https://medium.com/hackernoon/use-of-ml-and-ai-in-blackhat-hacking-737a621e4694>.
- Chi, H., Jones, E., Chatmon, C. and Evans, D. (2009). "Design and implementation of digital forensics labs", *Proceedings of the 12 iasted international conference on computers and advanced technology in education*.
- Chisholm, J. A. (2015). *Analysis on the perceived usefulness of hands-on virtual labs in cybersecurity classes*. Ph.D. dissertation. Colorado Technical University.
- Choi, K.-S. (2015). *Cybercriminology and Digital Investigation*. Lfb Scholarly Pub LLC.
- Choi, K.-S., Lee, C. S. and Louderback, E. R. (2020). "Historical Evolutions of Cybercrime: From Computer Crime to Cybercrime", *The Palgrave Handbook of International Cybercrime and Cyberdeviance* (pp. 27-43). Palgrave Macmillan.
- Du, W. and Wang, R. (2008). "SEED: a Suite of Instructional Laboratories for Computer Security Education (Extended Version)", *The ACM J Educational Resour Comput (JERIC)*, Vol 8 No. 1, Article 3.
- Gonzalez-Manzano, L. and De Fuentes, J.M. (2019). "Design recommendations for online cybersecurity courses", *Computers & Security*, Vol 80, pp. 238-256.
- Heemskerk, I., Kuiper, E. and Meijer, J. (2014). "Interactive whiteboard and virtual learning environment combined: effects on mathematics education", *Journal of Computer Assisted Learning*, Vol 30 No. 5, pp. 465-478.
- Kessler, G. C. (2007a). "Exercises and methodologies teaching hands-on cyberforensics skills online", In D. Edgar-Nevill (Ed.), *Proceedings of CFET 2007: 1st International Conference on Cybercrime Forensics Education and Training*, September 6-7, Canterbury Christ Church University, Canterbury, UK.
- Kessler, G. C. (2007b). Online education in computer and digital forensics: a case study, *Proceedings of the 40th Hawaii International Conference on System Sciences*.
- Liu, J. (2016). "Ten-Year Synthesis Review: A Baccalaureate Program in Computer Forensics", *SIGITE '16: Proceedings of the 17th Annual Conference on Information Technology Education*. September 2016, pp. 121-126.
- Liu, Z. and Wang, N. (2009). "Developing a Computer Forensics Program in Police Higher Education", *Proceedings of 2009 4th International Conference on Computer Science & Education*, pp. 1431-1436.
- Mison, A., Davies, G. and Eden, P. (2020). "The future direction of cybercrime and the difficulties of digital investigations: a rationale for a review of digital investigation specialist education", *ECCWS 2020 20th European conference on cyber warfare and security*. T. Eze, L. Speakman and C. Onwubiko (Eds.), pp. 249-257.
- Morawo, A., Sun, C. and Lowden, M. (2020). "Enhancing engagement during live virtual learning using interactive quizzes", *Medical Education*, Vol 54 No. 12, p. 1188.
- Olfman, L. and Bostrom, R.P. (1991). "End-user software training: an experimental comparison of methods to enhance motivation", *Information Systems Journal*, Vol 1, No. 4, pp. 249-266.
- Pedersen, M.K., Skyum, B., Heck, R., Müller, R., Bason, M., Lieberoth, A. and Sherson, J.F. (2016). "Virtual learning environment for interactive engagement with advanced quantum mechanics", *Physical Review Physics Education Research*, Vol 12, 013102.
- Schoonenboom, J. and Johnson, R.B. (2017). "How to construct a mixed methods research design", *Kolner Z Soz Sozpsychol*, Vol 69, 107-131.
- Stark-Wroblewski, K., Ahlering, R.F. and Brill, F.M. (2007). "Toward a more comprehensive approach to evaluating teaching effectiveness: supplementing student evaluations of teaching with pre-post learning measures", *Assessment & Evaluation in Higher Education*, Vol 32, No. 4, pp. 403-415.
- Ward, P. (2020). "Development of a Small Cybersecurity Program at a Community College", *Proceedings of the EDSIG Conference*, p. 4901.
- Wiedenbeck, S. and Zila, P.L. (1997). "Hands-on practice in learning to use software: a comparison of exercise, exploration, and combined formats", *ACM Transactions on Computer-Human Interaction*, Vol 4, No. 2, pp. 169-196.
- Yasinsac, A., Erbacher, R.F., Marks, D.G., Pollitt, M.M. and Sommer, P.M. (2003). "Computer Forensics Education", *IEEE Security & Privacy*, Vol 1, No. 4, pp. 15-23.
- Yaqoob, I., Ahmed, E., Rehman, M. S., Ahmed, A. I. A., Al-garadi, M. A., Imran, M., and Guizani, M. (2017). "The rise of ransomware and emerging security challenges in the Internet of Things", *Computer Networks*, Vol 129, pp. 444-458.
- Yuan, X., Williams, K., Yu, H., Chu, B., Rorer, A., Yang, L., Kizza, J. and Winters, K. (2014). "A Workshop on Teaching Information Assurance through Case Studies and Hands-on Experiences", *Proceedings of the 47th Hawaii International Conference on System Science (HICSS 2014)*.
- Zhang, X. and Choo, K.-K. R. (2020). *Digital Forensic Education: An Experiential Learning Approach*. Springer.

Protecting Networks with Intelligent Diodes

Jason Dahlstrom¹ and Stephen Taylor²

¹Web Sensing LLC, Lebanon NH, USA

²Thayer School of Engineering, Dartmouth College, Hanover NH, USA

Jason.Dahlstrom@websensing.com

Stephen.Taylor@dartmouth.edu

Abstract: This paper explores the utility, practical nuances, performance characteristics, and attendant security risks associated with intelligent *Diodes* -- network appliances that regulate traffic flow by validating formats and mission payloads within the security perimeter afforded by a single Field Programmable Gate Array (FPGA). Diodes operate in the middle-ground between networks that are fully air-gapped -- i.e., completely disconnected from the Internet -- and those that are fully connected but require complex boundary defenses and security administration. As such, they provide an all-hardware, real-time alternative for protecting military vehicles and sensitive networks. Diodes are particularly useful in four core settings: when industrial plant need be connected to cloud-analytic services, such as Google-Analytics, for the purpose of process optimization; for supporting the lifecycle of military vehicles through Condition-based Maintenance; for preventing information bleed when sensor feeds must be consumed inside sensitive networks; and finally, for the reliable distributed replication of large files and databases. The devices may operate directly on protocol-specific traffic headers to dispatch or block traffic. Alternatively, they may be customized to validate file and traffic formats. This is achieved using an automated circuit design workflow that builds a hardware parsing plugin from a formal grammar and embeds it into the FPGA. This results in custom hardware that intelligently operates on mission specific data.

Keywords: data diode, google-analytics, condition-based maintenance, FPGA, cloud connectivity, data leak prevention

1. Distribution

This research is supported by the Defense Advanced Research Projects Agency (DARPA) under contract W31P4Q-20-C-0033. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views of the Department of Defense or the U.S. Government. The paper is released with the following distribution determination:

Distribution A: (Approved for Public Release, Distribution Unlimited).

2. Introduction

Many organizations handle sensitive data and files relating to military missions, trade secrets, intellectual property, and/or private personal data (NBAR, 2021). Traditionally, these organizations have been well advised to implement an “airgap” that physically disconnects computers containing sensitive information from any connection to the Internet to protect against theft. Unfortunately, airgaps come with a substantial cost in productivity and maintenance as they assume that the relevant staff, with the expertise to handle network isolation, are co-located. Airgaps are also increasingly impractical given the need to connect critical systems, such as industrial plant, to cloud-based analytic platforms, such as Google Analytics, for the purpose of process optimization, or to support condition-based vehicle maintenance (Adams 2012, Carter 2013, Shanthakumaran 2010).

A technology alternative is the High Assurance Guard (HAG) (Heckman, Schell, and Reed 2012) often used in military cross-domain applications to share information between networks at different classification levels. These devices operate at the boundary between network segments to control and filter the flow of information; they often involve sophisticated virtual machine architectures and software stacks that must be supported by highly skilled network professionals with cumbersome incident reporting and investigative processes.

This paper considers an alternative *Diode* appliance, one instance of which is shown in Figure 1 (Dahlstrom and Taylor 2018, Aug 2019, Oct 2019, 2020, 2021). This device is marketed by Web Sensing LLC (websensing.com), a New Hampshire small business focussed on network security research. Internally, the device consists of a single custom circuit board (shown right). By design, the board employs a *single* FPGA chip (outlined) laying directly between the two leftmost Gigabit Ethernet (GigE) interfaces. The appliance thereby forms a “bump-in-the-wire” between any pair of the network segments, with the FPGA forming the bridge for all communication. Consequently, it can monitor and interact with all systems attached to its interfaces and can act to store, analyze,

and either forward or drop the traffic between them. It is important to note that outside the boundary of the FPGA, all the other circuitry is purely concerned with power and signal conditioning.



Figure 1: Diode Appliance

As an *all-hardware* appliance, the Diode offers several key security advantages: All sensitive data -- encryption keys, buses, and algorithmic functionality -- is strictly hidden within the security perimeter afforded by the FPGA chip-boundary, thereby mitigating reverse engineering in the event that a Diode is lost or stolen; No software is present in the device, thereby mitigating malicious implants and zero-day attacks and preventing its use as a steppingstone in sophisticated network attacks (Kushner 2013). An additional GigE connection (shown on the right) provides an isolated *out-of-band channel* that can be used to monitor the device or adapt its protection profile. Extensive anti-tamper techniques have been developed to enhance its resilience including the elimination of a JTAG interface and the ability to actively destroy circuits.

All the circuits resident in the Diode are developed using a rapid prototyping technology termed High Level Synthesis (HLS). This process is outlined in Figure 2 which contrasts software (blue), hardware (green) and HLS (brown) development pipelines.

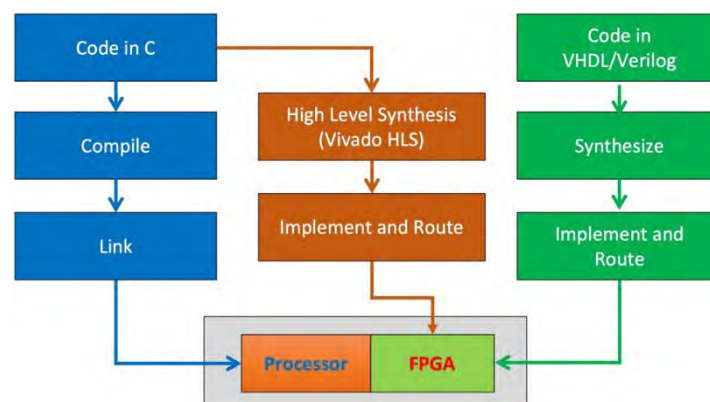


Figure 2: High-Level Synthesis

When generating software to run on a conventional processor, we generally develop code in some high-level language such as C, C++, or System-C, then compile each module independently, link the resulting object files, and bootstrap the resulting binary image on the processor. Conversely, when developing for FPGA's, we design a circuit in a Hardware Definition Language (HDL) such as VHDL or Verilog, then *synthesize* it to produce a set of gates implementing the circuit definition, then implement and route the gates with standard FPGA circuit cells comprised of Flip-Flops, Lookup-tables, and blocks of Random-Access Memory. The result is a *bit-file* used to configure the FPGA circuit resources.

High-Level Synthesis provides an alternative route to configure an FPGA: The circuit operation is designed in software, then *automatically* converted into a synthesized circuit, which is then implemented and routed for the FPGA. The advantage of this approach is that an algorithm can be designed, along with associated testbench code, by any systems programmer with little specialize knowledge of hardware. The working software can then be automatically transformed into a standalone *hardware block* which is validated under co-simulation using the original software testbench. The validated block forms a reusable *plugin* that can be directly integrated into any static circuit design in the FPGA. Alternatively, the block can be treated as *container*, compliant with the

Open Container Initiative (OCI) and akin to those used in Docker (c.f. docker.com). Using a technique known as *partial reconfiguration*, the FPGA can then be partitioned into compartments, or *slots*, linked together with industry standard Advanced eXtensible Interface (AXI) data-streams. Plugins can then be *dynamically* loaded into slots. To manage this process, a thin, all-hardware layer -- termed a *Nanomarshal* (Dahlstrom et al, Nov 2019) -- has been developed whose operation is akin to a bare-metal hypervisor (Barham, et al 2003). A variety of pin-compatible FPGAs, with differing resources, are available for use on Diode appliances. The Artix-7 is the smallest family that supports partial reconfiguration; however, smaller Spartan-7 devices are available for static designs.

3. Basic UDP Diode

The conceptual floorplan for the simplest Diode circuit is shown in Figure 3; it provides one-way transport of only User Datagram Protocol (UDP) traffic. There are two separate GigE data paths: from the *input* port (IN) to the *output* port (OUT) for outbound traffic (green), and vice-versa for inbound traffic (blue). Interposed within these paths are two different *packet inspection engine* (PIE) plugins. The outbound PIE (green) inspects the packet headers on all arriving Ethernet frames, passes only the UDP-frames, and drops any others. The inbound PIE (blue) forms an open circuit by dropping all the arriving frames, thereby forming the Diode; though this block is wired into the printed circuit board, there are no connections within the PIE circuit itself providing the required isolation.

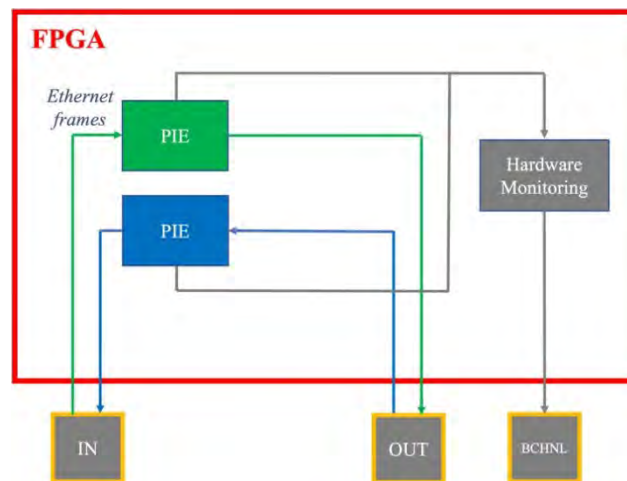


Figure 3: Diode Floorplan

The outbound PIE need only buffer a few bytes of the packet header to validate the protocol type; It can then either stream the remainder of the packet out of the device or drop it entirely. Consequently, the device achieves *gigabit rates with sub-microsecond latency*. Simple variations on this theme, that validate specific Ethernet MAC addresses, IP addresses, and/or UDP ports, can be added with negligible impact on performance.

Each PIE may accumulate statistical information concerning the number and type of packets being passed or dropped. This information can be made available on the backchannel periodically via a monitoring plugin that generates UDP or Simple Network Management Protocol Trap messages. These can be consumed by a standard Network Monitoring System (NMS), such as OpenNMS, Solarwinds, Nagios, Zabbix, etc. attached to the backchannel.

Since all inbound traffic is dropped, including Address Resolution Protocol (ARP) requests, it is necessary to manually configure ARP cache entries for devices attached to the Diode ports; this is a trivial exercise that can be achieved on Linux using the command:

```
sudo arp -s <IP_address> <MAC_address>
```

It is relatively trivial to add ARP validation to both PIE's and avoid the inconvenience, with a slight increase in risk and decrease in performance.

3.1 Use Case: Protecting High-value Assets and Condition-based Maintenance

The core use case for the UDP *Diode* is to protect a single *high-value asset* connected directly to monitoring or analytics services, such as Google Analytics, for the purpose of process optimization as illustrated in Figure 4. Generally, these assets are designated systems whose compromise or damage would seriously impact corporate operations. Examples of such systems associated with industrial plant include laser and machine cutters; production-line assembly plant; HVAC and electrical supply systems; blast-furnace control systems; refinery control systems; waste and water treatment control systems; databases of production patterns, intellectual property, and facility details. In the medical arena, high-value assets include life-support systems, medical imagers, radiology equipment, and patient databases.

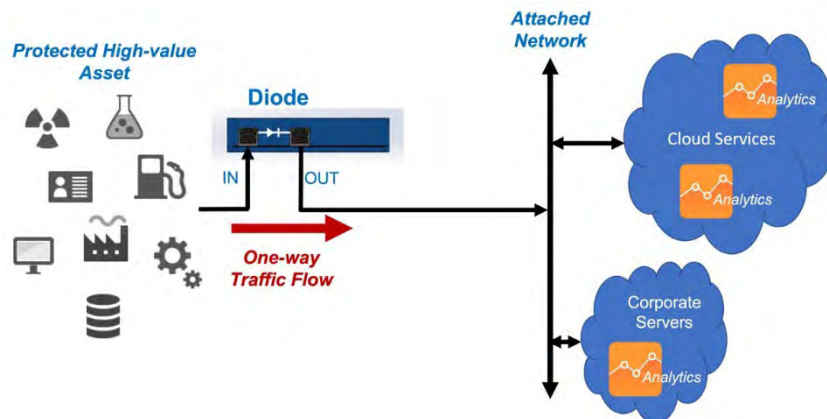


Figure 4: Protecting High-Value Assets

Since *Diodes* enforce one-way traffic in hardware, they provide isolation with a high degree of assurance against software attacks. To use a Diode in this manner it is co-located physically near the protected asset. The Ethernet cable at the asset is unplugged and attached to the Diode output; a short patch-cable is then connected from the asset to the Diode input. In this configuration, the asset may communicate status and performance data directly to analytics engines either at locally connected corporate servers, or alternatively, to cloud services located on the Internet. Since only one-way traffic flow is allowed, no traffic may pass from potentially compromised hosts downstream of the Diode either from within the organization or the Internet. This configuration is particularly useful when only periodic updates are required, or for periodically updating information between geographically distributed sites, such as corporate offices or hospitals.

The increasing prevalence of Condition-Based Maintenance (CBM) makes this configuration valuable for monitoring heavy vehicles, when combined in-line with wireless connectivity at the output port (OUT). The central idea is that vehicles are continuously monitored for indications of wear and failure with sensor data periodically uploaded to remote analytic servers. The analytics combine sensor data with historical component failure data and predictive models to determine when parts are likely to fail or need replacement, allowing replacements to be provided just-in-time, thereby reducing downtime, delays, and costs.

3.2 Adding Encryption with Open-Source Software

It is straight-forward to reliably transmit encrypted, signed files across a UDP Diode, then validate and decrypt them at a receiver. The Diode is directly connected between two servers: an inside or *client* server, located on a protected network behind the Diode, that serves up encrypted files and an outside server, that receives and decrypts the files as illustrated in Figure 5.

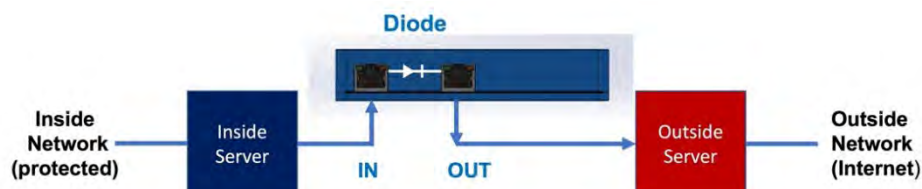


Figure 5: Encrypting Transfers

Each server must be equipped with the OpenSSL and UDPcast tools. OpenSSL provides a rich set of cryptographic tools for data privacy, authenticity, and integrity checking. OpenSSL's asymmetric encryption capabilities are also used to "sign" an encrypted file. The process generates a signature file that can be used to verify the

authenticity of the sender as well as to validate the integrity of the file's contents. At the sender, it is used to encrypt and sign the file on the client (file sender) and then to verify and decrypt the file on the outside server (file recipient). UDPcast provides software capable of transferring data across a Diode. It provides data redundancy and data ordering in the one-way data stream necessary because the data receiver cannot alert the data sender of corrupted, missing, or out-of-order data.

Initially, a public/private key pair is generated for the *outside* server, for example, with the commands:

```
openssl genrsa -aes256 -out server.private.key 8912
openssl rsa -in server.private.key -pubout -out server.public.key
```

The *server.private.key* is stored on the *outside* server and kept secret. The *server.public.key* is shared with anyone wishing to encrypt data intended for the outside server. As such, a copy of the *server.public.key* should be placed on the *inside* server. Similarly, a key pair for signing is generated on the *inside* server using the same commands – this time generating *client.private.key* and *client.public.key* files. The *client.private.key* is stored on the *client* server and kept secret. The *client.public.key* is shared with anyone wishing to verify authenticity and integrity attributes of files signed by the client. As such, a copy of the *client.public.key* should be placed on the *outside* server. The sharing process for a file (e.g. *message.txt*) is then straightforward using the following steps:

1. Encrypt the file to transmit using the *outside* server's public key:
`openssl rsautl -encrypt -pubin -inkey server.public.key -in message.txt -out encrypted.file`
2. Sign the file to transmit using the *inside* server's private key:
`openssl dgst -sha512 -sign client.private.key -out encrypted.file.sha512 encrypted.file`
3. Combine the encrypted file and its digest into a compressed archive bundle:
`tar -zcvf bundle.tgz encrypted.file encrypted.file.sha512`
4. Start the udpcast server code on the *outside* server:
`udp-receiver --nokbd --receive-timeout 60 -f received.bundle.tgz`
5. Send the bundle on the *inside* server using the udpcast client code:
`udp-sender --async --max-bitrate 5m --fec 8x8 --file bundle.tgz`
6. Uncompress and unpack the compressed archive file on the *outside* server:
`tar -zxvf received.bundle.tgz`
7. Verify the authenticity of the bundle's sender and the bundle's file integrity:
`openssl dgst -sha512 -verify client.public.key -signature encrypted.file.sha512 encrypted.file`
8. Decrypt the encrypted file:
`openssl rsautl -decrypt -inkey server.private.key -in encrypted.file -out message.txt`

This entire process can easily be automated with simple shell scripts for Diode users.

Although the two servers could be anywhere, to simplify the process we have developed a complete *Gateway* appliance, shown in Figure 6. This system combines the Diode with two Next Unit of Computing (NUC) computers (top left and right) that act as the designated servers, running either Windows or Linux. This system also includes isolated power supplies, associated with the inside and outside networks, to protect against power-line exploitation across the Diode. Each server has two network interfaces: one to the Diode, and the other to either the inside or outside networks. Since the network connections to the Diode are direct, they are extremely low-loss making transport reliable.



Figure 6: Gateway Appliance

3.3 Use Case: Low-value Assets and Database Replication

Many industries maintain large files and databases associated with their core industrial processes to track and optimize production or billing, and for historical purposes. Some organizations also require the ability to securely copy them between geographically dispersed locations. Thus, some applications may have stringent *reliability* concerns that require the guarantee that all data transitioning a Diode is successfully received.

Turnstile technology is a patented hardware mechanism that provides *reliable* one-way transfers without violating the one-way data flow guarantees provided by a Diode (Dahlstrom and Taylor, 2021). The technology is fully contained within the FPGA circuit design shown in Figure 3 and thus can be used with either Diode or Gateway appliances. The essential idea is that the sending computer or server wraps its transmissions with size and sequencing information. The receiver generates an acknowledgement upon successful completion of a transfer, that is intercepted on the inbound link at the Diode. Hardware within the Diode internally generates a completely isolated response to the sender indicating the successful completion of the transfer. This occurs *without* transferring any information from the output to the input of the Diode, ensuring that no network-borne exploits can transition the Diode.

There are often non-essential or support assets within an organization that serve ancillary engineering and administrative services, such as storage and manipulation of intellectual property, payroll, or personal data. Elements of each individual service constantly exchange information and the interposition of individual Diodes, as shown in Figure 4, would cause unacceptable interference in normal operations. Unfortunately, these lower-value systems can serve as an easy target if unprotected, allowing theft of intellectual property, or simply acting as a foothold within the organization for information gathering or launching network attacks later. In such instances, it is preferable to protect independent segments of the Local Area Network (LAN) that contain a set of related systems or databases as shown in Figure 7: Here a single Diode protects an entire LAN segment -- with all its workstations, systems, databases, and servers -- allowing them to inter-operate freely as a single entity. Turnstile technology allows reliable controlled transfers between such segments, akin to air-gapping them.

Typically, LAN's are organized with a collection of rack mounted network switches and routers placed in routing closets. Individual segments are connected into the switching infrastructure through standard *patch-panels* allowing for flexible reconfiguration. The Diode and Gateway appliances can thus be easily inserted between segments, simply by moving cables.

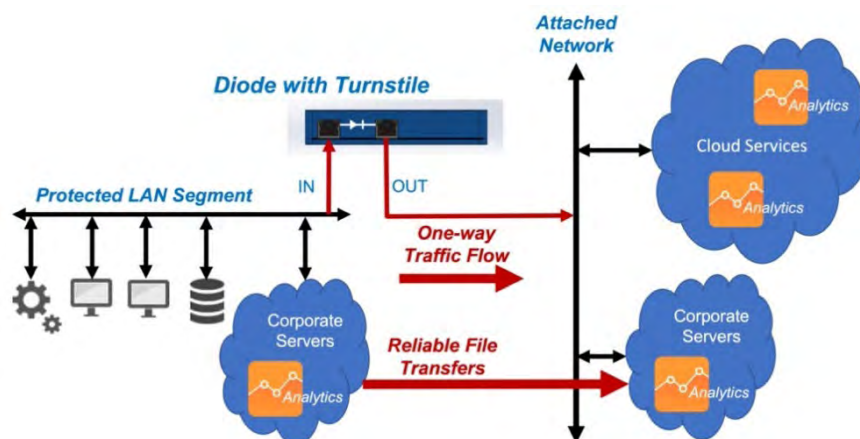


Figure 7: Reliable Transfers

4. Building Diode Intelligence

Figure 8 refines Figure 4 and provides the conceptual floorplan of an intelligent Diode, illustrating typical flows and plugin blocks that can be inserted within the FPGA either statically or dynamically. As before, there are two basic data paths: from the *input* port (IN) to the *output* port (OUT) for outgoing traffic, and vice-versa for incoming traffic.

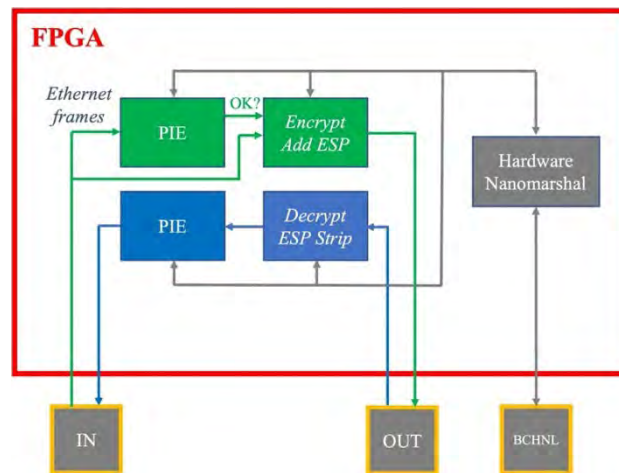


Figure 8: Intelligent Diode Floorplan

The green path in Figure 3 illustrates a typical outbound flow that concurrently inspects, encrypts, and encapsulates Ethernet frames completely in hardware. If the packet inspection algorithm determines that a frame is valid -- in terms of its protocol headers and payload content -- it signals (c.f. OK) the downstream blocks causing them to transmit the encrypted frame within an encapsulation payload; otherwise, the encrypted frame is discarded. Similarly, a typical inbound flow, shown in blue, might strip encapsulation and decrypt a frame, then validate the inbound frame and either pass or drop it.

Each of these blocks can be replaced by the degenerative case of a *wire*, that passes all traffic, or an *open circuit*, that drops all traffic. Consequently, a wide variety of appliances can be realized with the same basic structure. For example, the simple User Datagram Protocol (UDP) Diode discussed previously is composed of a green path that checks the IP protocol header to ensure each frame is UDP; the encryption/encapsulation block is replaced by a wire, and the inbound path by an open circuit. Thus, only valid UDP packets pass from input to output, and all inbound traffic is dropped. Alternatively, a Transport Control Protocol (TCP) Diode involves an inbound path that passes only valid connection requests and data packets, and an outbound path that allows only valid TCP flow synchronization packets.

The third (rightmost) GigE port providing the backchannel in Figure 8, serves to provide a direct connection to Nanomarshal hardware allowing plugins to be dynamically installed within the dataflow of the device; alternatively, it may be populated by a simple monitoring block used Figure 4.

4.1 Hardware Encryption Performance

The appliance shown in Figure 8 may leverage *any* encryption and/or encapsulation scheme rendered into a plugin via HLS. By default, an AES-256 plugin, certified using the NIST Cryptographic Algorithm Verification Program (CAVP) is available, with the standard Encapsulating Security Payload (ESP).

It is possible to render the AES algorithm into a wide variety of alternative implementations using pragmas that inform HLS how to layout and route the attendance circuits. This results in a trade-space that allows circuit performance -- bandwidth, maximum clock frequency, and latency -- to be successively traded for FPGA resources -- Block RAM, Flip-flops, and Lookup-Tables.

Benchmark studies led to the discovery of a “sweet spot” in this trade-space that achieves close to Gigabit line speed -- 0.82 Gb/sec -- with only ~2.5% LUT utilization on an Artix 200T FPGA device. Note that generally two instances of the encryption block are required: one for encryption and one for concurrent decryption, with similar resource requirements. Though it is possible to achieve 1.1Gb/sec -- exceeding line speed -- this comes at heavy cost in LUT resources: ~20% of the Artix 200T FPGA per instance. Unfortunately, the AES Cyclic Block Chaining (CBC) mode of operation has a data dependency that limits pipelining. However, Counter mode relieves this dependency and yields performance approaching 10Gb/sec.

4.2 General Packet Inspection and its Performance

The typical action of a Diode on detecting a valid message is to allow it to pass; conversely, its action on detecting invalid data is to drop the message, thereby mitigating potential exploitation, but may also involve generating

an alert on the backchannel. Complex validation tasks, beyond the protocol header checks outlined in Section 3, might involve sophisticated tests to ensure that the messages adhere safely to industry standard protocols, that message fields contain only legitimate values, and/or that message content is valid within the context of a particular *mission*.

In general, *parsing* is the process of validating an input stream of symbols by understanding their format (syntax) and meaning (semantics). For example, compilers such as GCC use a parser to validate that a computer program, written in some programming language such as C/C++/Java/Fortran is written correctly (i.e. is syntactically valid), and to understand its structure (i.e. its semantics) for the purpose of machine code generation and optimization. Parsers are tools that apply a collection of formal *grammar rules*, defining some input language, to determine if the input adheres to the rules. For example, the following 3-rule grammar G defines a language in which a stream of symbols is valid only if it begins with the character 'a', ends with 'c', and contains one or more intervening 'b' characters:

G : 'a' Bs 'c' ;

Bs: 'b' | Bs 'b' ;

The "or" symbol | designates an alternative definition for the rule defining "Bs". This grammar accepts as valid the input streams *abc*, *abbc* and *abbbc* etc, but rejects any other stream, e.g. *a*, *ac*, *aaa*, *ccc*, *adx*, *abbbx*, etc. Individual characters such as 'a' are *terminal symbols* that must be present in the input stream; all other symbols are non-terminals representing intermediate structural elements. For binary grammars, hexadecimal terminal values can also be used (e.g., '\xFF' represents a single byte value corresponding to 255 in decimal).

Parser generators are tools that take a grammar as input and automatically generate a program that implements the associated parser. The most mature and widely used generator is Bison which accepts two primary classes of grammar: Generalized Look-Ahead (GLR) and the more restrictive Left-to-Right Look Ahead (LALR). Both classes of grammar are expressed in Backus-Naur Form (BNF), used above to define the grammar G. Under the DARPA SafeDocs program, new tools are being developed based on formal methods. One of the most mature is the Hammer *combinator* library which provides a collection of well-defined base parsers and methods to combine them to build more complex parsers. The resulting parsers are provably correct by construction. The Hammer library provides a collection of backends that allow different classes of grammar to be implemented, including GLR and LALR.

Though GLR grammars are more general, LALR grammars are sufficient for validating a wide variety of protocol and file formats, where the input is in either ASCII and/or binary. These grammars can be realized with a generic *push-down automaton* – a finite state machine employing a single stack to store symbols while parsing the input stream. The state machine relies on two core operations *shift* – involving saving a symbol from the input onto the stack and *reduce* – involving the application of a grammar rule to detect a structure in the input and reduce the symbols on the stack. The state-machine is generic and common to all grammars, however, the order in which shift and reduce operations are applied to the input is based on a collection of *parsing tables* derived from the grammar, usually referred to as Action/Goto tables (Aho, Sethi, and Ullman, 1988). These tables map the current state of the automaton, to a next state based on the symbol read from the input stream.

A completely automated process has been developed that takes an LALR grammar, such as G above -- written in either Bison or Hammer -- and produces a validated parser plugin that can be inserted into a Diode as a packet inspection engine (PIE). This process involves the automatic generation of an optimized set of parsing tables, their combination with generic automaton code, and a translation process based on HLS. The primary use of FPGA resources in the resulting parsers is BRAM used to store the parsing tables. For concreteness, consider the full JSON grammar (c.f. json.org). In Bison, it consists of approximately 325 rules and its corresponding state machine has 689 states. The associated implementation requires 1.46Mbytes of BRAM resources to implement, which corresponds to ~2% of the BRAM resource on an Artix 200T device; a negligible percentage of other resources is also required.

Parsing is inherently a byte-by-byte process that builds knowledge of the structure of the input stream seen up to a particular point. In consequence, the time to parse a message or file is directly proportional to the length of the input. For the above JSON grammar, the initial prototype has a processing time of approximately 45 clock cycles per byte; at 125 MHz the time required to process an Ethernet frame (1600 bytes) is 575 microseconds. Though the link may operate at Gigabit rates, when parsing the bandwidth drops -- by virtue of the need to

store, parse, and forward individual Ethernet frames sequentially -- generally varying between 2.25 and 3.5 Mb/sec. All our optimization efforts thus far have focussed on reducing FPGA resource requirements, to ensure that the parsers can fit into the FPGA's, rather than latency and bandwidth. Currently, the parsers constructed automatically are only able to validate syntax. For semantic checks, they are augmented with hand-coded C-functions prior to HLS, a process we expect to mature over time.

4.3 Use Case: Preventing Network Bleed

Generally, the perception of a Diode is that it protects assets from inbound attacks. However, it is important to understand that there is a reciprocal value when a Diode is configured in the reverse direction. Figure 9 shows this alternative with a Diode connected such that its input (IN) is connected to the outside network, while its output (OUT) is connected to the protected network. This configuration has the objective of allowing sensor feeds and database transfers from untrusted networks into the protected network, while ensuring that *no data can exit the protected network*, i.e., that no valuable information can escape, or bleed out of it.

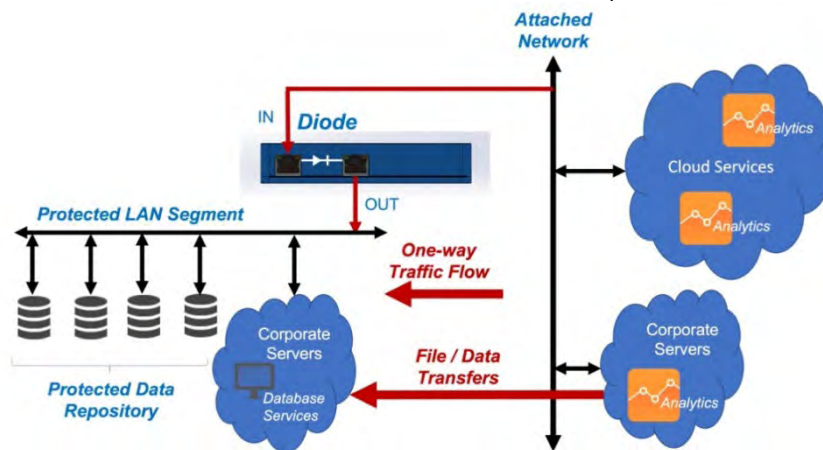


Figure 9: Preventing Information Bleed

Clearly the value of timely data collection must be weighed against the attendant risks associated with this approach. However, the generalized Diode offers the opportunity to insert highly specialized mission- and time-dependent validation into data streams to mitigate these risks.

5. In Conclusion

This paper has presented a methodology for building a wide variety of intelligent Diode appliances with their attendant use cases. The methodology is based on a collection of hardware plugins that can be statically configured or dynamically installed, to adapt to a perceived threat level, via an out-of-band channel. At one extreme the appliances use protocol header information to validate traffic in real-time. At the other extreme, sophisticated intelligence is built into them using automation techniques. This intelligence is capable of carefully checking every byte in a transfer to validate detailed mission-specific payloads. The use of multiple Diodes within a corporate environment allows network administrators to carefully control and shape the available attack surface to protect valuable assets. This provides precise control of what machines are accessible, from where, and what data is allowed to flow, in which direction between them.

6. Acknowledgements

The authors wish to thank and acknowledge the work of Stephen Padnos, who designed the Diode hardware, and Jim Brock who designed the automation workflows.

References

- Adams, C. (2012) "HUMS Technology", Aviation Today (aviationtoday.com/2012/05/01/hums-technology).
- Aho, A.V., Sethi, R., Ullman, J.D. (1988) *Compilers*, Addison Wesley.
- Barham P, et al (2003) "Xen and the Art of Virtualization", In: Proceedings of the nineteenth ACM symposium on Operating systems principles, pp 164–177.
- Carter, M.C. (2003) "Post Implementation CBM Benefit Analysis", Annual Conference of the Prognostics and Health Management Society.
- Dahlstrom, J., and Taylor, S. (2018) "System-on-Chip Data Security Appliance and Methods of Operating the Same", U.S. Patent 10,148,761.

- Dahlstrom, J., and Taylor, S. (Aug 2019) "System-on-Chip Data Security Appliance and Methods of Operating the Same", U.S. Patent 10,389,817.
- Dahlstrom, J., and Taylor, S. (Oct 2019) "Endpoints for Performing Distributed Sensing and Control and Methods of Operating the Same", US Patent 10,440,121.
- Dahlstrom, J., et al, (Nov 2019) "Hardening Containers for Cross-Domain Applications", In Proceeding of MILCOM 2019, Norfolk, VA, USA, pp. 1-6.
- Dahlstrom, J., and Taylor, S. (2020) "System-On-Chip Data Security Appliance Encryption Device and Methods of Operating the Same" U.S. Patent 10,616,344.
- Dahlstrom, J., and Taylor, S. (2021) "Hardware Turnstile", U.S. Patent 10,938,913.
- Goldberg R.P. (1974) "A survey of virtual machine research", In: Proceedings of Computer, 7th edn., pp 34–45.
- Heckman, M.R., Schell, R.R., and Reed, E.E. (Dec 2012) "Toward formal evaluation of a high-assurance guard", In: Proceedings of the 6th Layered Assurance Workshop (LAW) 2012, December 2012, Orlando, FL. pp. 25-29.
- Kushner, D. (2013) "The Real Story of STUXNET", IEEE Spectrum.
- NBAR (2021) -- The National Bureau of Asian Research, The IP Commission Report.
- Shanthakumaran, P. (2010) "Usage Based Fatigue Damage Calculation for AH-64 Apache Dynamic Components", The American Helicopter Society 66th Annual Forum, Phoenix, Arizona.

Can Attrition Theory Provide Insight for Cyber Warfare?

Stephen Defibaugh and Donna Schaeffer

Marymount University, Arlington Virginia, United States of America

Stephen_Defibaugh@marymount.edu

dschaeff@marymount.edu

Abstract: This paper explores the notion that cyber-adversaries can use classic attrition tactics to cause weakness to address follow-on attacks. We conducted a grounded theory study that reviewed historic literature to identify parallels between past attrition tactics and cyber warfare. From historical examples, we see the possibility of an adversary conducting an asymmetric campaign by flooding the adversary with false-positive attacks in order to have them drain resources. For a modern perspective, we interviewed subject-matter experts from a US military command. Thematic analysis demonstrates a link between attrition and cyber-maneuver warfare. One significant finding is that most subject-matter experts agreed a culture of compliance, which encourages a full resources response to security events given full resources, can reduce the ability to maneuver appropriately and takes away from the focus on critical mission functions that cyber security is actually in place to protect. Other common themes that surfaced include that some interviewees believed their organizations were not prepared for cyber war nor are they resourced adequately to respond to a state of cyber war. Issues that need further study are the need to compare and correlate telemetry and metrics of incident responses and better tracking of the dollar-cost value of incident response and cyber tactics.

Keywords: cyber, attrition, compliance, cyber warfare

1. Introduction

With the emergence of cyber space as a domain of future warfare, it remains a contested territory that has many competitors present maintaining an equal number of independent objectives and goals. As shown through history, with the emergence of each new domain, such as sea, air, and space, we see potential maneuvers for advantage and domination of each unexplored area (Defibaugh, 2021). Often times, these competitions bring diverse groups and individuals into conflict. Accepting this, it is not surprising to see maneuvering extend into the cyber realm (Applegate, 2012). One of the challenges is the popularity in the mentality of deterring or preventing the single attack or multiple attacks aimed from a single actor. In many cases, this is a vague and multi-headed beast even in the best of circumstances. There are often many solutions for the different types of problems, and very few instances of a clear-cut answer that will address a cyber security need. In the face of this particular problem many leaders turn to compliance-based policy and guidance that provide generic steps and mitigations (Defibaugh, 2021). These generic steps and mitigations are often used as the only measurement for success in today's current environment. In this mentality of regulatory or policy compliance, organizations seemingly desire to spend more resources in an attempt to remain compliant in the eyes of regulatory agencies and the media, regardless of impact or defense against the likelihood of adversary engagement. These provide short-term mitigations to specific issues but do not always scale up to consider the longer-term, higher perspective.

Many decisions by cyber security professionals or organizational leadership are often made as responses to intelligence that is gathered regarding threats to an organization's weaknesses. These threats can come from a wide spectrum of possible actors ranging in skill and motivation; from basic scripting and building skills to advanced adversaries looking to establish a near-permanent foothold within systems. The advanced adversaries are more willing to play a long game and may not always use the most sophisticated methods when simpler tactics will achieve the same goal. The advanced use of tactics, techniques, and procedures (TTPs) are often like a fingerprint that are used in their attribution. Increasingly, we see signs that advanced persistent threats (APTs) are provided resources by, or are directly controlled by, nation-states in an effort to either disrupt or gain another nation or organization's resources (Defibaugh, 2021). In 2006, elements of the United States Air Force Cyber Command and the Department of Defense began to incorporate the concept of cyber warfare into their operational doctrines (Butler 2013).

However, as time has progressed, those involved in these seeming cyber warfare campaigns have witnessed a regular increase in threats and attackers with no real definition of what the indicators of compromise (IOCs) or the attributes of a cyber war might even look like. In an environment where the use of advanced TTPs in a cyber war could mean using tactics similar to kinetic warfare, we may begin to see the use of cyber as a means for attrition warfare instead of the highly technical attacks we are used to seeing publicized in the media. Large,

complex organizations may even be exposing themselves to such attacks by attrition by reacting to each individual threat or false attack in a very predictable manner.

2. Examining the Strategic Impact of a Compliance & Response Mentality

Despite increased activity by threat actors and recognition of cyber threats in the global media, studies have not fully explored the concept that reacting to every security incident with full resources could be used as a method to attack an organization, possibly denying or degrading the targeted organization from accomplishing its mission or goals (Defibaugh, 2021). The second topic of concern is that there are no available indicators of compromise that identify when malicious cyber activities go beyond small-scale attacks and into the realm of cyber warfare. The amount of qualitative and quantitative research on a cyber-attack through attrition as a viable attack method has not been fully realized. There are many resources that early on identified cyberspace as the next domain of warfare (Harris 2018). Other works focus on the possibility of future multi-domain warfare where cyber wars will be fought alongside more conventional kinetic combat. There are a surprising number of early works documented from 2006 through 2012 that suggest that there are bridges between kinetic and cyber warfare. Nevertheless, none of the materials identified in the early research has attempted to study any direct use of attrition tactics as viable methods of cyber warfare. Even so, some organizations appear very willing to proceed without any consideration or exploration into the idea of regular drain on resources being weaponized against them. Any thought of exposure to a consistent resource drain, despite long-term impact that might prevent them from adequately responding to a single or multiple adversary attackers, is entirely suppressed.

This article addresses two issues. The first issue is the exploratory examination of classic attrition tactics used by advanced cyber adversaries to exploit the current compliance driven and reactionary posture of United States-based organizations in order to weaken them as a part of a coordinated cyber warfare campaign. Second, qualitative research was conducted to determine if experts in the cyber security field have considered the concept of attrition as a tactic to conduct cyber warfare (Defibaugh, 2021). In order to examine this gap in the research, an exploratory qualitative methodology, using an exploratory grounded-theory approach, examined historical examples of warfare where attrition was used to degrade or deny adversarial advantages. The preconceived notions of attritional conflict leading into World War I will be examined along with literature that reviews maneuvering and relationships between attack and defense costs. This research study focuses on the exploration of the concept where cyber attrition tactics as a component of a cyber war campaign to deny or degrade US-based organizations as it is executed at a nation-state and organizational level.

3. Assumptions, Limitations, and Scope

A limitation of this research is that it is exploratory in nature and will only serve to essentially expose and explore a potential area in cyber security research. It will not be able to provide mitigations or corrections if the research questions are answered. The scope of this research will consist of cyber war and cyber warfare that would be executed at the nation-state level. As stated by Scott Applegate (2012), cyberspace appears to be more of a contested territorial domain that includes competitors from state and non-state domains, organized groups, proxies, and individuals. For the purposes of this research, electronic warfare and cyber terrorism will not be the focus. As this is an exploratory research case, the methodology was utilized from similar exploratory studies (Defibaugh, 2021). An assumption that is present during the course of this research is that cyber security experts that occupy similar positions with similar qualifications are qualified to maintain informed, professional opinions that are grounded in logic and reason.

4. Reviewing Conflicts and Cyber Security

William Philpott (2015) defines attrition as, “the cumulative exhaustion of the enemy’s fighting capacity” (Philpott, 2014). When we consider wars of attrition we often think about a number of different scenarios ranging from a siege of a fortified installation to people in trenches unable to advance or retreat. There are many examples of attrition warfare in human history. However, if we consider a series of events that involves multiple nation-state actors that include warfare beyond simple shooting, and that has been well documented, then the events that take place in the World War I provide an excellent area to examine. During World War I, the German military term for the strategy of attrition under State Siege Law was *Ermattungsstrategie* (Philpott, 2014).

Attritional strategy during conflict is effectively predicated on destroying or annihilating adversarial forces faster than the adversary could effectively replace them (Philpott, 2014). However, in 1915, and even in some cases before that time, a shift in strategic thinking began to take place on both sides of the conflict. Before the war

was started in earnest, the German general staff outlined a plan, which concluded that a protracted fight might end up ruining Germany economically. They faced opposition on multiple fronts, and if one opponent could be annihilated, then a second opponent might be worn down through an attritional struggle (Philpott, 2014). We note that in history, their choice had an impact on post-World War I German economics. Military strategy was decidedly amateurish. The failure of the leaders on both sides was often attributed to the differentiations between tactics, operations, and strategies, while completely ignoring the huge influence of resources and the constraints of logistics (Philpott, 2014). Presently, a worrisome parallel in the field of cyber security and cyber warfare emerges. If decision makers and subject matter experts are unable to take a step away from the immediacy of cyber security decision loops, then we may be resigning ourselves to repeat the same mistake and confuse what we think is strategy with tactical operations. When an adversary is able to invade the decision loop of another and force them, from a strategic level, to operate on a small, tactical scale, then the defending side is fighting a losing battle. This was a concept that Lloyd George, the British Prime Minister from 1916 to 1922, fundamentally failed to grasp. When both sides have millions of men in reserve and an efficient method of communication, most strategic initiatives will result in a stalemate (Lloyd George, 1982). This plays into the cyber realm very well if we consider the Internet as a force multiplier. In a modern sense, we may have numerous enemies that can exist in the cyber realm, as they are easily created and the Internet provides them with a nearly unlimited means of efficient communication. The French strategy, in July 1915, was to overthrow the Germans with simultaneous attacks that would not necessarily annihilate them outright, but would more-so overwhelm them on multiple fronts (Prete, 1915). The principle strategy of Erik von Falkenhayn, the second Chief of the German General Staff, was aimed at breaking the cohesion and will between Germany's adversaries. He planned to do this individually and collectively by destabilizing their united front. The thought process was that by engaging in battles on multiple fronts and inflicting heavy losses, this might encourage capitulation (Philpott, 2014).

5. Strategy During Cyber Maneuvering

We see in the previous section that conventional warfare not only shares a potential complementary relationship, but that cyber warfare is able to incorporate elements of conventional warfare using the Internet and information technology as a force multiplier. Next, we will examine how maneuvering in the physical domains has already begun to translate into maneuvering during cyber operations. According to Scott Applegate (2012), though conventional and cyber warfare differ over the amount of time that activities take place and their relation to varying strategies, the ultimate principle of maneuvering during a conflict, either cyber or conventional, has often been an important, if not deciding factor. If we look at historical examples of new domains and territories, it is unsurprising that cyberspace is a new domain where dominance is sought. Exploration of domains, such as land and sea, became the battlefields where new maneuvers and tactics were developed. Later, the same occurred for air and space (Applegate, 2012). As we consider cyberspace as a war-fighting domain, it is expected that it will become a contested territory vied for by numerous competitors, which may include state and non-state actors, organized groups, and even private individuals (Applegate, 2012).

The US military describes the concept of maneuvering as, "The disposition of forces to conduct operations by securing positional advantages before and or during combat operations" (Joint Publication 3-0, 2011). In historical examples, maneuvering would look like armies attempting to capture or defend strategic locations, governments may attempt to sequester or impinge on another nation's ability to engage in maneuvers or warfare through economic or strategic means, or any method or process that would give one actor an edge over another. In cyber terms, this may more resemble methods or processes employed to attack or defend information, which would constitute as the same since it provides an edge to one side over another (Applegate, 2012). In present conflicts, the targets of value are now focused on the capture of critical computing resources, industry and military secrets, and the infiltration of critical civilian and military infrastructure (Applegate, 2012). These targets, just like the Gallipoli Peninsula and Vimy Ridge in World War I, are critical areas of leverage, and the capture or denial of these would give one actor the edge over the other. This is an example that illustrates a clear parallel between the underpinned paradigms of conventional warfare and cyber warfare.

Colonel John Boyd described maneuvering as, "Operating inside of an adversary's decision loops and penetrating his moral-mental-physical well-being in order to destroy him from inside and remove his will to resist (Boyd, 1986). In this context, we start to see an application of an overall strategy forming. The established goal of attritional warfare is to exhaust the enemy's fighting strength and will to resist. Operating inside an enemies' decision cycle and eroding their will to continue fighting seems to be an effective way of conducting warfare. This type of attack is even more effective than those only in the physical domain, when the attack leverages the Internet as a force multiplier, which is both effective and very inexpensive to conduct. When we frame strategies

in this manner, then we can start to better understand the idea of cyber maneuvering as an application of force. The purpose of this application of force may be to capture, disrupt, deny, degrade, destroy, or manipulate computing and information resources in order to achieve a position of advantage over other competitors (Applegate, 2012). According to Parks and Duggan, from their *Principles of Cyber Warfare*, “Maneuvering in cyberspace involved the application of force to specify points of attack or defense” (Duggan, 2011). Furthermore, according to Scott Applegate (2012), the idea of cyber maneuvering as an element of warfare has come along far enough for us to identify defining characteristics. Though research is being conducted into types of cyber warfare maneuvering, the extent of which these maneuvers are employed, which might define a clear act of cyber warfare, has not yet been internationally established (Mali, 2018). Willson & Tomhave define a weapon of cyber warfare as an information technology system purpose built to damage the infrastructure or operations of another IT system (2012). This angle does not take into account the possibility of a cyber warfare weapon being used to alter or direct the behavior of an adversary to a desired point or conclusion.

6. The Discovered Link between Attack and Defense Costs

From the literature, we can see that there is a clear link between conventional warfare and cyber warfare. In some circumstances, we even see the integration of cyber war alongside conventional tactics, or as a prelude, in order to soften an adversary prior to a significant strike. Policy makers and executive leaders, although recognizing the value and risk of attritional warfare, have already begun to move away from this and consider different approaches (Defibaugh, 2021). Peter Munson (2007) wrote that strategic policy makers and leaders must be able to more accurately assess the available resources and thus set strategies that are appropriate to limited power, without overreaching. From the literature, we can see that there is a clear link between conventional warfare and cyber warfare. In some circumstances, we even see the integration of cyber war alongside conventional tactics, or as a prelude, in order to soften an adversary prior to a significant strike. As far as defining cyber capabilities as weapons, it is often equated that the term weapon implies an offensive connotation, whereas some weapons may actually be a passive force that is used to degrade or eliminate an adversary’s will to resist (Mali, 2018). The Tallinn Manual defines a weapon of cyber warfare as something that is capable of causing injury, damage, death, or destruction (Schmitt, 2013). However, this representation does not take into account passive weapon systems designed to degrade or interfere with an adversary (Mali, 2018). Though Mali does further define weapons of cyber warfare, he does not apply them to a stratagem. Applegate takes this further by establishing that not only are maneuvers still relevant but, using the Internet as a force multiplier, they are effective and potentially devastating if misinterpreted. However, at present, we lack framing mechanism to understand what tactics and maneuvers are available in order to employ a strategy of cyber attrition against an adversary.

From this, we are not only able to define cyber war as the ability to steal, degrade, interrupt, or otherwise interfere with an adversary’s information systems, but also the maneuvering of force in the cyber realm in order to gain a significant advantage. We are also able to better understand the idea of specific maneuvers that would enable these activities and interpret them, not as individual events, but on the greater sum of their activities towards an end game (Defibaugh, 2021). Unlike the definition of cyber warfare and cyber weapons put forth by Rid & McBurney, the cyber attacks themselves may not be effective, but have the ultimate goal of changing an adversary’s thought process entirely and thus manipulating their behavior favorably towards the adversary’s goals (2012). We also see that not only is it possible to interfere with an adversary and manipulate their decision-making process, but that it has even been executed in several public and well documented international events. Having previously established attritional warfare as a relevant strategy in the modern era, it is a viable strategy in an era of asymmetric conflicts using the Internet as a tremendous force multiplier. We begin to see possible exploitation, positional, and influencing maneuvers that are readily available to attack our more classic mentality and effectively wear us down through resources or will attrition, until we ultimately lose the collective ability to continue resisting. According to the Internet Security Alliance Board of Directors, “An estimated \$1 trillion was lost in the US in 2008 through cyber attacks” (Williams, 2009). Even the response required merely to figure out who perpetrated a cyber attack requires expensive resources and talent in the form of specialized teams with specific knowledge and experience. With the sharp increase in frequency of cyber attacks, we also see a proportional increase in spending (Jensen, 2009, 2010). The concern that is becoming evident is that US-based organizations, with their policies of strict compliance, are reacting in proportion to each individual cyber event or incident and bringing the full scope of resources to bear. This is a problem when you start to pull back and look at the larger picture of sustainment in the cyber realm. To secure complex systems, it takes thousands or tens of thousands of trained personnel and billions of dollars involved in monitoring, maintaining, and defending

complex systems of systems (Lynn, 2010). Members of the United States Senate have even stated that the federal government has spent more on cyber security for federal agencies than it has on the gross domestic product (GDP) of North Korea (US Congress Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Federal Financial Management Government Information Federal Services and International Security, 2010). This is very troubling because if we look at historic examples from William Philpott (2015) and the First World War I, we see that economies of adversary nations were clear and definite targets for attrition style warfare. It can be noted in several cases where the cost of maintaining armed forces, infrastructure, or the inability to adequately respond prevented the defending side from responding in any meaningful way. In several historical circumstances, to include Vimy Ridge, it was evident that nations were able to maneuver themselves in ways to exploit their adversary's circumstances to ultimately cripple them and erode their will to resist until they ultimately surrendered. From the perspective of an adversary who is able to cheaply and effectively use the Internet as a force multiplier, it is advantageous for them to allow the US and NATO to continue chasing every single incident without any regard to potential maneuvering or orchestration by an adversary seeking to execute cyber maneuvers in support of a larger intentional campaign (Defibaugh, 2021). An adversary seeking to gain an advantage through an influencing maneuvering may be attempting to support an attrition campaign designed for friendly organizations to expend precious resources chasing shadows and threats that were never truly intended to have an impact on their own. Only by looking at the aggregate maneuvers do we start to really see the potential jeopardy that a reactionary posture seeking to blindly enforce compliance, provides.

7. Researching Attrition Theory for Cyber Warfare

In order to examine the modern topic of attrition theory in the context of cyber warfare, the general concept of attrition warfare had to be established as a grounded and accepted fact. In order to achieve this, attrition through history was reviewed using a methodical gathering and analysis of data. The concept of attrition as a strategy and tactic in conventional warfare is well grounded in the historical literature. Using both online and print sources, major wars and conflicts can be found utilizing keywords searches. The grounded theory research supports the research questions by underpinning circumstances where wearing an adversary down through indirect combat, economical, or social means was a more advantageous avenue than outright head-on conflict. Having confirmed that attrition is a viable attack strategy, a comparison of cyber warfare strategies can be undertaken to show distinct circumstances where attritional warfare would have a potential impact utilizing the combat domain of cyber. In order to examine the modern topic of attrition as a strategy in cyber warfare, the general concept of attrition warfare had to be established as a grounded and accepted fact. In order to achieve this, attrition through history was reviewed using a methodical gathering and analysis of data. The concept of attrition as a strategy and tactic in conventional warfare is well grounded in the historical literature. Using both online and print sources, major wars and conflicts can be found utilizing keywords searches. The grounded theory research supports the research questions by underpinning circumstances where wearing an adversary down through indirect combat, economical, or social means was a more advantageous avenue than outright head-on conflict. Having confirmed that attrition is a viable attack strategy, a comparison of cyber warfare strategies can be undertaken to show distinct circumstances where attritional warfare would have a potential impact utilizing the combat domain of cyber.

The interview protocol was developed to support the research method for gathering data to determine if economic factors could potentially be used to engage an organization in attrition warfare. The first component was the number of individuals interviewed in order to gather significant data. Based on expert recommendations from researchers and experts in the same field, the number of interviews was a minimum of seven individuals. The individual interviewees all belonged to the same parent organization, but also belonged to different sub-organizations. In each circumstance, the interviewee was not a direct subordinate, nor were they supervised by the research team. The interview subjects were each of the same pay grades and had the same minimum qualifications as established by the organization. For this study, since all of the subject matter experts were required to have the same qualifications to hold their position and grade, the interviewees' race, gender, and other influencing factors were deemed irrelevant to this area of research.

The interview protocol that was utilized was designed to limit confirmation bias and sample bias. In order to support this, the subject matter expert had to be within a certain organizational grade level and have equivalent cyber security roles and responsibilities. As shown in previous research, cyber security is a relatively new field in comparison with others and, therefore, is still exploratory (Defibaugh, 2021). The interview method was built upon the interview protocol to support the primary research topics, specifically the question of whether financial

and thematic data could demonstrate a relationship to cyber war by visualizing attrition activities against an organization. The first step in the interview method was to develop the questions that would permit a semi-structured format with open-ended questions that would elicit unbiased, clear answers. The desired amount of questions for an hour-long interview was about 11 to 15 questions. (Defibaugh, 2021)

8. Struggling to See Beyond Compliance

According to qualitative interviews conducted at a U.S. based organization, two common themes were brought to light from the data aggregated from the following examples (Defibaugh 2021):

1. Some organizations in this study were not prepared for cyber war.
2. The focus of cyber security is entirely myopically centered on a compliance-based mentality, according to a majority of participants in this study.

According to research performed by Defibaugh, it was established that US-based organizations, using a strict-policy of compliance, are reacting proportionally to individual cyber events and incidents (2021). This becomes a significant problem when the defenders have the perception of having to be right in every encounter with the adversary, whereas they only have to be right once. As the topic of compliance furthered, some participants expressed concern of compliance as the seemingly only benchmark for cyber warfare readiness. One of the interviewees from Defibaugh's study (2021) offered that,

As a matter of fact, some of the programs that are being put in place actually decreased performance within our activity. And so then we asked specific questions about, hey, if we have to perform our mission based on the restrictive or based on the configuration that you put in front of us, how do we perform that mission? And the answer is, 'Well, you have to work around that.' So then, we ask when we get interviewed by the people who are chartered to go put this protective measure in place at the top level domain, and we start telling the things we do, it doesn't appear to me that they understand our mission or the technologies we're using (p. 77).

This is an interesting perspective in that it does not simply mean to be done with compliance audits, moreover, that the total sum effect to include larger impact on strategy must be accounted for. The participant wondered, in the same vein as attrition warfare, if cyber security professionals could audit themselves into submission (Defibaugh, 2021):

I think primarily most of our organizations are still focused on compliance. I think they're more concerned with just making sure that they're compliant, that they're meeting the requirement, and they stop at that point. And I think the perspective has to change to where we look at that. That compliance that we're looking at is the baseline, not the end goal. The end goal is to take it beyond that and start looking at how you can actively start making changes that aren't part of that requirement. It's training and perspective, the perspective from management in general has to change to where we're not just looking to be compliant. Compliance is a part of the picture, but it's not the goal. And I think the second part is that the only way people are going to truly start to understand when they see something anomalous and understand and recognize it as a problem, is if they run tests in their own environment (p. 78).

Another participant felt very strongly that compliance activities had become overblown and,

So, what happens when you get compliance-based compliance, then it is validated with an audit and audit comes in? They make sure that all of those compliance pieces are in place. So what happens is you basically shift all of your focus from providing your day-to-day production support to making sure that you're compliant, and somewhere in the middle of that it gets lost. The fact that the really important thing is that you're monitoring the traffic levels and what's happening on your network, so that if you are attacked, you can recognize it (p. 78).

This is an interesting perspective in that it does not simply mean to be done with compliance audits, moreover, that the total sum effect to include larger impact on strategy must be accounted for. Qualitative research provides contextual information that does not always have readily available quantitative measures with which to perform a completely rounded analysis.

Table 1: Data Collection Categorization developed from (Defibaugh, 2021)

Data Value	Data Type	Expected Value Collected
Typical Number of Events Experienced in 12 months?	Interval	No
Average amount of funding expended per month on Incident Response for 12 months?	Interval	No
Percentage of incidents determined to be false positive across 12 months?	Interval	No
Are organizations focused on spending resources appropriately on cyber attack?	Nominal	Yes
Does a compliance driven mentality weaken an ability to respond to cyber threats?	Nominal	Yes

As seen in Table 1, common themes extracted from the nominal interview data can be used to demonstrate the gap in quantitative data when the metadata is compared (Defibaugh, 2021). Each of the interviewees were able to provide nominal data based on earned experience and tacit knowledge. However, there was a significant lack of available interval data to perform a meaningful quantitative validation on the contextual knowledge that was collected.

9. Conclusions and Future Areas of Research

One of the overarching themes that was identified in the research conducted by Stephen Defibaugh was that in many cases, the blind pursuit of compliance takes away from the focus on things which may truly be more critical. In circumstances where the behaviors of a target are known, then the compliance based approach begins to show weaknesses as an overall strategy. Compliance sets the bare minimum standard and baselines systems in order to provide a basic profile to system security. Yet, when adversaries understand that stimuli can be created to manipulate behaviors and decisions, then we start to see where compulsive behavior and reactions can become very dangerous. Further topics of research would center on further exploration of qualitative themes gathered from cyber security experts that have a unique view of both policy, operational execution, and resource expenditure. The qualitative, exploratory research that was conducted by Defibaugh demonstrates the need for further quantitative and mix-methods studies that examine the possibility of utilizing attritional styles of attack as a strategy for conducting cyber warfare. Understanding the amount that is spend on compliance as well as incident response must be balanced against strategic perspectives and pragmatic worldviews. It is imperative that the cyber security field understand the cost-value impact of its currently adopted methodologies and industry practices by developing the proper metrics and telemetry that can identify patterns at a strategic level.

References

- Applegate, S. D. (2012, June). The principle of maneuver in cyber operations. In *2012 4th International Conference on Cyber Conflict (CYCON 2012)*, (pp. 1-13). Retrieved from <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6243974&isnumber=6243954>
- Boyd, J. R. (1986). Patterns of Conflict. Project on Government Oversight, Defense, and National Interest - John Boyd Compendium. J. R., 1986. Retrieved from <http://dnipogo.org/john-r-boyd>
- Butler, S.C (2013). Refocusing on Cyber Warfare Thought. *Air & Space Power Journal*, 27(1 (Jan/Feb)), 44-57. Retrieved from <http://proxymu.wrlc.org/login?url=https://search-proquest-com.proxymu.wrlc.org/docview/1318929576?accountid=27975>
- Defibaugh, S. J. (2021). *Attrition as a Practical Attack Methodology and Tactic for Cyber Warfare* (Doctoral dissertation, Marymount University).
- Harris, M. A. (2018). Preparing for Multidomain Warfare: Lessons from Space/Cyber Operations. *Air & Space Power Journal*, 32(3), 45. Retrieved from <http://proxymu.wrlc.org/login?url=https://search-proquest-com.proxymu.wrlc.org/docview/2099884315?accountid=27975>
- Lloyd George, D. L. (2017). *War Memoirs of David Lloyd George*. Vol. 1 Pickle Partners Publishing.
- Mali, P. (2019). Defining Cyber Weapon in Context of Technology and Law. In *Multigenerational Online Behavior and Media Use: Concepts, Methodologies, Tools, and Applications* (pp. 1119-1132). IGI Global.
- Munson, P. (2007). *The Return to Attrition: Warfare in the Late Nation-State Era*. NAVAL POSTGRADUATE SCHOOL MONTEREY CA CENTER FOR CONTEMPORARY CONFLICT..
- Parks, R. C., & Duggan, D. P. (2011). Joint Publication 3-0 Joint Chiefs of Staff, United States Department of Defense, Washington, D. C., **2011**, pp. 111-27.
- Parks, R. C., & Duggan, D. P. (2011). Principles of Cyber Warfare. *IEEE Security & Privacy*, 9(5), 30-35.
- Philpott, W. J. (2015). *War of Attrition: Fighting the First World War*. Abrams.

- Prete, R. A. (1997). The Franco-British strategic conflict on the Western Front and the Calais conference of July 6, 1915. *World wars and contemporary conflicts*, 17-49.
- Rid, T., & McBurney, P. (2012). Cyber-Weapons. *The RUSI Journal*, 157(1), 6–13. doi:10.1080/03 071847.2012.664354
- Schmitt, M. N. (2013). *Tallinn Manual on the International Law Applicable to Cyber Warfare*. Cambridge University Press. doi: 10.1017/CBO9781139169288
- US Congress Senate Committee on Homeland Security and Governmental Affairs. Subcommittee on Federal Financial Management Government Information Federal Services and International Security. (2010). *More Security, Less Waste: What Makes Sense for our Federal Cyber Defense: Hearing Before the Federal Financial Management, Government Information, Federal Services, and International Security Subcommittee of the Committee on Homeland Security and Governmental Affairs, United States Senate of the One Hundred Eleventh Congress, First Session, October 29, 2009*. US Printing Office.
- Williams, M. (2009). Cyberspace May Be Locale of Next War. *Federal Times* (Dec. 7, 2009).
- Willson, D., & Tomhave, B. (2012). *Legal & Ethical Considerations of Offensive Cyber-Operations*. RSA Conference 2012. Retrieved from https://www.rsaconference.com/writable/presentations/file_upload/star-304.pdf

A Novel DevSecOps Model for Robust Security in an MQTT Internet of Things

Manasa Ekoramaradhya and Christina Thorpe
Technological University Dublin, Blanchardstown, Dublin, Ireland
manasa151685@gmail.com
christina.thorpe@tudublin.ie

Abstract: Message Queuing Telemetry Transport (MQTT) is a standard messaging protocol for the Internet of Things (IoT). Among the various communication protocols used in IoT, MQTT stands unique because of its multiple advantages such as being efficient and light weight, reliability in message delivery and scalability to millions of things. However, the fact that the data privacy of the MQTT messages can be compromised while the data is in transit poses risks to the security mechanism. Attack scenarios related to MQTT have exposed multiple risks and vulnerabilities such as thousands of MQTT brokers being accessible over the default port, data privacy, authentication, data integrity, port obscurity, and botnet over MQTT. These risks and vulnerabilities undermine security mechanism which results in compromised IoT systems. Development Security and Operations (DevSecOps) aims at integrating security at every phase of the IoT lifecycle with enhanced automation, tools, and a process for determining security vulnerabilities at every stage. This results in a rapid and cost-effective IoT system which is enabled by proactive security mechanisms, threat prediction, threat detection, and alerting mechanisms. The aim of this work is to build a DevSecOps pipeline utilizing open source MQTT servers and brokers. A comparative study was performed to identify the risk posture provided by the DevSecOps pipeline across MQTT ports offering different combinations of security mechanisms. Firstly, threat modelling was conducted wherein the IoT system was analyzed at an architectural level from an attacker's perspective and appropriate risk mitigation and defense mechanisms were accommodated into the design. The IoT system was then subjected to rigorous static and dynamic analysis followed by vulnerability scanning and third component checks. Penetration test cases and controls are automated to check threats and vulnerabilities like escalation of privileges, denial of service, spoofing, information disclosure, and repudiation. An alerting mechanism is also integrated into the system to monitor risks and vulnerabilities. Our proposed DevSecOps models achieves standard maturity in security systems with earlier threat prediction and detection.

Keywords: Message Queuing Telemetry Transport (MQTT), Internet of Things (IoT), Development Security and Operations (DevSecOps), Cybersecurity

1. Introduction

An increasing number of devices connected to the Internet of Things (IoT) is contributing to increased cyber threats and cybercrimes across the world. For example, we now commonly see connected appliances, smart home security systems, autonomous farming equipment, wearable health monitors, smart manufacturing equipment, etc. These connected devices pose a threat to cyber security for several reasons (Hassija 2019) (Meneghello 2019): firstly, there is an abundance of data transmitted and stored by these IoT systems; secondly, the IoT systems are hyper connected to both virtual and physical environments; thirdly, there is massive complexity in IoT environments due to the volume and diversity of devices and connections; and finally, many of these devices have been manufactured with security as an afterthought. It is reported that cybercrime alone cost \$1 Trillion USD in 2020 (McAfee 2020) and is estimated to reach over 10 Trillion annually by 2025, driving the need for cybersecurity solutions or systems for IoT. This is especially important given the critical role IoT will play in applications such Industry 4.0 and Intelligent Transport (Okano 2017).

Examining IoT systems at an architectural level reveals that it uses many protocols such as CoAP, MQTT, XMPP, RESTFUL Services, 6LoWPAN and many more (Sharma 2018), however, the MQTT protocol has gained widespread use for machine-to-machine communication in bandwidth-constrained environments where compact data transport is a strict requirement. The multiple advantages of MQTT protocol makes it a very attractive option for IoT devices. These include minimal resource requirements, lightweight, efficiency, message persistence and the ability to support bi-directional communications (Mishra 2020).

Despite their popularity, MQTT and IoT devices came under wide scrutiny when Palo Alto Networks reported that 98% of all IoT device traffic exposed private and confidential information on to the network as it was unencrypted; and 57% of these devices were susceptible to medium or high severity attacks which attracted the hackers (Palo Alto 2020). Extensive research on MQTT security has proved that multiple brokers were discovered by Shodan (Harsha 2018), which we also confirmed. Since MQTT by default does not involve any encryption standards, it poses a high risk in terms of data privacy. MQTT authentication can be compromised by extracting

X.509 certificates or by sniffing techniques. Posts to an MQTT node authenticates at a broker, there is no specific process to determine which publishers or subscribers can publish or receive messages, posing serious authorization issues (Syaiful 2017) (Palo Alto 2020). When penetration testing was done on open MQTT systems, multiple risks and vulnerabilities were revealed related to elevation of privileges, denial of service and many more.

In the past decade, traditional development cycles have become less common with many moving towards a more lean and agile Development and Operations (DevOps) cycle. DevOps means end-to-end automation in software development and delivery (Ebert 2016) (Zhu 2016). While this new paradigm presents many benefits in the form of agility and efficiency, it does pose a risk for security as many of the robust security processes are no longer integrated as the deployment is placed in the control of the developers (Fox 2020) (Perera 2016). Absence of an efficient threat detection and alerting mechanism while applying a DevOps approach in MQTT application development leads to the risks with high severities skipping the focus of the development team.

Development, Security and Operations (DevSecOps) introduces enhanced automation and aims at security integration at every stage of software development lifecycle (Kenyon 2021). Implementation of this method in IoT systems helps in identifying security risks from the initial design stage until the live stage. It also integrates threat detection and alerting mechanisms for advanced monitoring. This includes security integration starting from static analysis and source code reviews, vulnerability scanning, scanning for components with known vulnerabilities and third-party components. Dynamic analysis and extensive penetration testing are done to reveal any vulnerabilities that might be potentially exploited. However, while there has been work on developing a DevSecOps (Myrbakken 2017) models for Cloud applications (Rajasekharaiah 2021) (Kumar 2021) and IoT systems (Bahaa 2021) (Wolf 2021) there is no work in the literature on DevSecOps for MQTT specifically.

Using a DevSecOps model, the entire architecture of the IoT system which uses the MQTT protocol will undergo extensive threat modelling to determine the possible risks in the system. The overall risk posture will be determined as a result of all these DevSecOps activities and the integrated alerting mechanisms will send notifications of these results. In this paper, we will compare the results of DevSecOps methodology which was implemented across three different port services offered by the opensource MQTT broker, Mosquitto. Our results show that our model achieves standard maturity in security systems with earlier threat prediction and detection.

The remainder of the paper is organised as follows: Section 2 discusses the technical background information and the related work pertinent to the research presented in this paper. Section 3 defines the methodology defined and implemented in this research. Section 4 presents the results of the experimentation conducted. Finally, Section 5 concludes this paper and presents items for future work.

2. Background and Related Work

Though MQTT is widely used across most of the IoT systems, the security vulnerabilities and the risks associated with this protocol poses severe challenges. DevSecOps deals with proactively identifying and monitoring these vulnerabilities. The below sections detail out the working and security challenges associated with MQTT and the DevSecOps methodology.

2.1 MQTT

MQTT is an OASIS standard protocol and hence the specifications is provided by OASIS MQTT Technical Committee¹. **MQTT v3.1.1** contained a number of threats that had to be considered, such as : devices were vulnerable to compromise; data at rest in clients and servers were potentially accessible; protocol behaviors could have side effects (e.g. “timing attacks”); Denial of Service (DoS) attacks; communications could be intercepted, altered, re-routed or disclosed; and spoofed Control Packets could be injected. The implementations had to consider solutions such as: authentication of users and devices; authorization of access to server resources; integrity of MQTT Control Packets and application data contained therein; privacy of MQTT Control Packets and application data contained therein. **MQTT v5.0** has enhanced security features - both MQTT client and server must provide: authentication, authorization, and secure communications options.

¹ <http://docs.oasis-open.org/mqtt/mqtt/>

MQTT is a TCP/IP based protocol which requires an established TCP connection before communication. A three-phase handshake is employed; in first phase, the TCP connection is established; in the second phase, MQTT connections and data transfer takes place; the third phase terminates the established TCP connection. The reliability of the MQTT messages depends on the QoS (Quality of Service) levels used. The selection depends on the type and importance of the MQTT messages/topics. In level 0, the transmission is unreliable, the message is delivered without duplication and no acknowledgement is required. In level 1 (default), the message is delivered at least once without duplication and an acknowledgement is required. In level 2, the message is delivered without duplication.

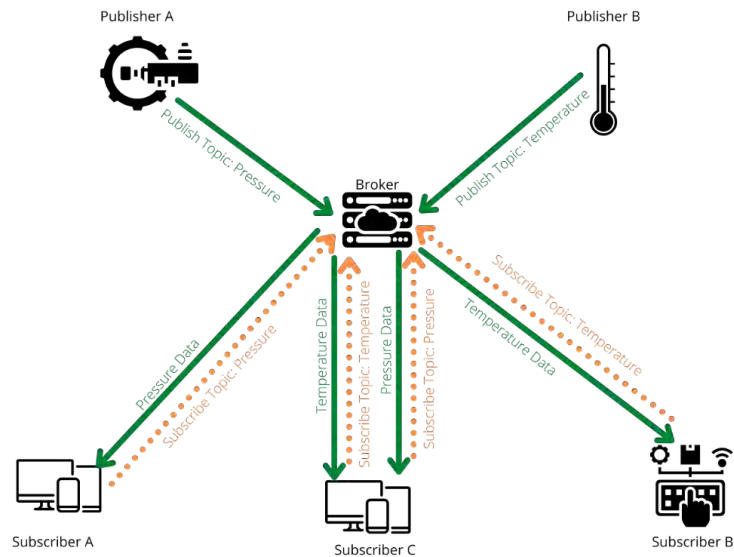


Figure 1: Typical MQTT publish-subscribe scenario

2.2 Typical MQTT publish-subscribe model

MQTT is an application layer protocol which relies on TCP and works on publisher-subscriber model (Mishra 2020). Figure 1 illustrates a typical MQTT communication wherein the centralized broker intermediates the communications between the publisher and the subscriber nodes. *Publisher A* sends the data related to the topic 'Pressure' and the *Subscriber A* receives this data by subscribing to this topic via the *Broker*. A subscriber can subscribe to multiple topics (as illustrated by *Subscriber C*) and any node in the network can act as publishers and subscribers. The connection between the broker and the client nodes will always stay open and the transmitted packet size is relatively small. These factors speed up the MQTT communications when compared to other protocols. MQTT protocol consumes low power and is efficient for communicating shorter messages, which supports high scalability.

2.3 Security risks and vulnerabilities in MQTT

Though MQTT is a highly resourceful protocol it poses several risks in terms of security. Previous research work has shown that thousands of MQTT brokers were openly available on port 1883 and were indexed by Shodan (Syaiful 2017). This research also highlighted that the MQTT brokers with client connection code 0 did not have any client authentication mechanism which enable anonymous clients to either publish a message or subscribe to any topic on that network. Since MQTT does not provide any encryption protocol by default the privacy of the data in transit can be easily compromised by capturing the packets with the help of tools such as Wireshark or Burpsuite. After the KeepAlive time expires the publisher must resend the connect packet which contains the username and password. Capturing this packet will eventually compromise the MQTT authentication mechanism. After the capture of the packet traffic the attacker can easily modify the data which leads to the compromise in data integrity (Syaiful 2017). Furthermore, MQTT is susceptible to fuzzing and a lot of vulnerabilities were detected by fuzzing techniques (Casteur 2020). MQTT has shown severe risks when it comes to replay attacks and Man-In-The-Middle attacks (Chen 2020). MQTT has also posed severe challenges with respect to confidentiality, access control, and Denial of Service (Hintaw 2019). Moreover, since these vulnerabilities are mostly discovered from a hacker's perspective or while they are in production, fixing the vulnerabilities becomes a challenging task. This points to the increased need of the system which will integrate

security at every phase of the IoT system development while using MQTT and drives a more proactive approach in threat hunting and alerting which can be fulfilled by a DevSecOps methodology.

2.4 Development, Security and Operations (DevSecOps)

DevSecOps brings in the idea of enhanced automation and security integration in the Software Development Lifecycle (SDLC). It aims at implementing security at each phase starting from the initial design phase till the production and maintenance phase (Kenyon 2021). This is more efficient and a cost-effective method since finding the vulnerabilities at the early stages of the software development Lifecycle are easy to be fixed than the ones found in live or production. This introduces a proactive way of finding the vulnerabilities and it also accelerates the patching of security vulnerabilities. Nearly half of all software development organizations are now relying on this model to implement a proactive security methodology. The Gartner Hype Cycle for Agile and DevOps, 2020, indicates that DevSecOps is in the early stages of mainstream adoption. Gartner quotes a 20-50% market penetration among DevSecOps' target audience. Gartner projects that DevSecOps will reach mainstream adoption within 2-5 years. Figure 2 illustrates the concept of DevSecOps and how it relates to the more well-known DevOps phases.

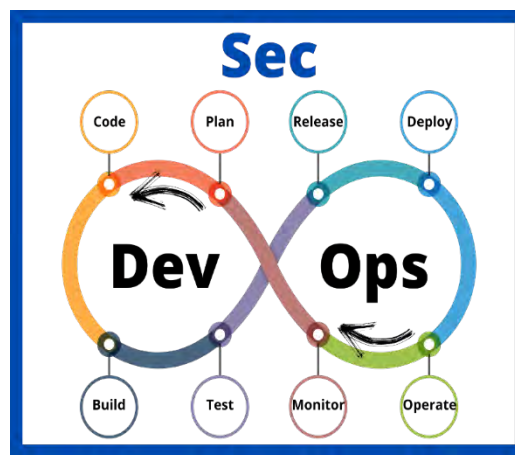


Figure 2: General DevSecOps Model

3. Methodology

Implementing DevSecOps for IoT systems and devices, especially where a protocol like MQTT is being used for communications, helps in identifying the security risks, vulnerabilities, and severities at every stage. Along with detection and identification, treating/controlling these risks can happen in a more efficient and cost-effective manner.

3.1 Implementation

Figure 3 illustrates the DevSecOps pipeline developed for an MQTT IoT system. In the following sections, we describe the DevSecOps activities that form the pipeline. Each of the three test scenarios were assessed using the sample pipeline to validate our model. The tools selected for each phase are popular tools used by security professionals.

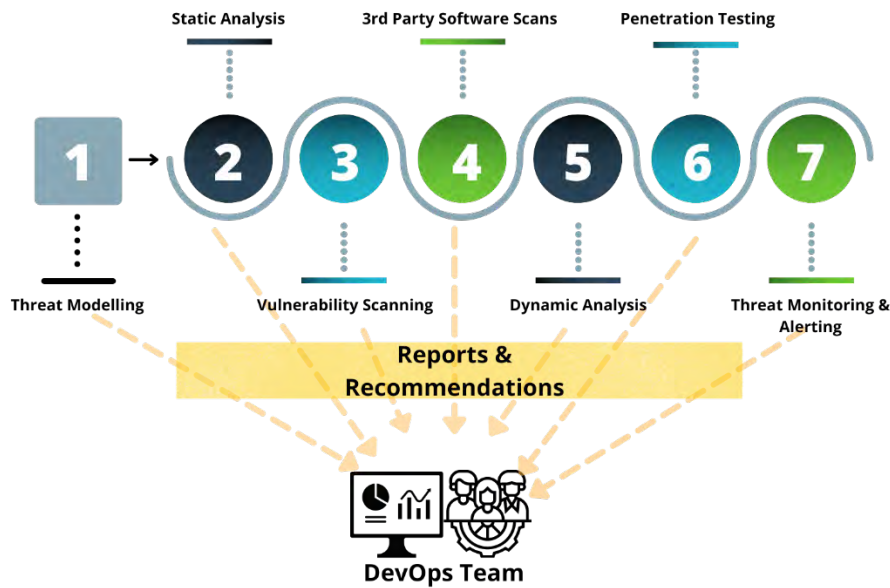


Figure 3: DevSecOps MQTT Pipeline

3.1.1 Threat Modelling

In this process, the architecture of the IoT model and the MQTT protocol were analyzed to identify and locate possible vulnerabilities, identifying probabilities and objectives for possible attacks and putting proper mitigation methods in place. This activity can be done manually, or to enhance the efficiency and accuracy, tools such as Microsoft Threat Modelling can be used. Many risks and vulnerabilities related to elevation of privileges, denial of service, information disclosure, spoofing, repudiation, device isolation was found in this stage. The Threat Modeling Tool is a core element of the Microsoft Security Development Lifecycle (SDL). It allows software architects to identify and mitigate potential security issues early when they are relatively easy and cost-effective to resolve. As a result, it greatly reduces the total cost of development. Also, the tool was designed with non-security experts in mind, making threat modeling easier for all developers by providing clear guidance on creating and analyzing threat models.

3.1.2 Static Analysis

Static Application Security Testing (SAST) is a white box testing method where the tester has access to the source code, framework, and the binary execution files. The vulnerabilities found at this phase are less expensive to fix. Tools used in this phase are CheckMarx and SonarQub. Checkmarx CxSAST is a highly accurate and flexible Static Code Analysis Tool that allows organizations to automatically scan un-compiled/un-built code and identify hundreds of security vulnerabilities in the most prevalent coding languages.

SonarQube is a Code Quality Assurance tool that collects and analyzes source code and provides reports for the code quality of your project. It combines static and dynamic analysis tools and enables quality to be measured continually over time.

3.1.3 Vulnerability Scanning

Web application vulnerabilities scanners help in scanning of the web applications from an outsider perspective to find common web application vulnerabilities like cross-site scripting, SQL injection, command injection and many more. In this phase the Nessus port scanner tool is used. Nessus is a remote security scanning tool, which scans a computer and raises an alert if it discovers any vulnerabilities that malicious hackers could use to gain access to any computer you have connected to a network. It does this by running over 1200 checks on a given computer, testing to see if any of these attacks could be used to break into the computer or otherwise harm it.

3.1.4 3rd Party Software Scans

The source code used to develop the devices or systems, or sometimes the protocols themselves, may have libraries which do not contain the latest code version, resulting in a lot of vulnerabilities. Tools such as BlackDuck and OWASP Dependency checker are used in this phase. Black Duck is a complete open-source management solution, which fully discovers all open source in your code. It can map components to known vulnerabilities

and identify license and component quality risks. You can use Black Duck to set and enforce open-source policies and integrate open-source management into your DevOps environment. Additionally, Black Duck monitors and alerts you when new threats are reported. Dependency-Check is a Software Composition Analysis (SCA) tool that attempts to detect publicly disclosed vulnerabilities contained within a project's dependencies. It does this by determining if there is a Common Platform Enumeration (CPE) identifier for a given dependency. If found, it will generate a report linking to the associated CVE entries

3.1.5 Dynamic Analysis

This is more of a black box testing, wherein the approach is like that of a hacker. Vulnerabilities here are discovered after the development phase and hence it is comparatively expensive to fix these. Tools such as OWASP Zap and HCL Appscan are run in this phase. Appscan runs automatic scans that explore and test web applications and includes one of the most powerful scanning engines in the world.

3.1.6 Penetration Testing

Penetration tests are cyber-attacks which are simulated against the system to identify the vulnerabilities which could potentially be exploited by malicious actors. It describes the intentional launching of simulated cyberattacks by "white hat" penetration testers using strategies and tools designed to access or exploit computer systems, networks, websites, and applications. Although the main objective of pen testing is to identify exploitable issues so that effective security controls can be implemented, security professionals can also use penetration testing techniques, along with specialized testing tools, to test the robustness of an organization's security policies, its regulatory compliance, its employees' security awareness, and the organization's ability to identify and respond to security issues and incidents such as unauthorized access, as they occur.

3.1.7 Threat Monitoring and Alerting

This is one of the important phases in DevSecOps. Several mature tools like Splunk can be integrated into the pipeline which will record the vulnerabilities, risks and sometimes even analyses the packets to find threats and block the source. When the behavior deviates, or a certain type of risk is found the alerting feature will automatically notify the user about the risks and vulnerabilities. It also keeps a count on the number of occurrences, source of the threat and uses machine learning algorithms to even predict the threats early.

3.2 Test environment

For this work, we used an IoT system with MQTTv5.0, configured to use QoS level 1 and other default parameters, as its base communication protocol for our DevSecOps implementation. Here, open-source servers and brokers from Mosquitto.org are used to set publisher and subscriber nodes and MQTT service provided across three different ports including:

- *Test Scenario 1 (Low Security):* port 1883 (MQTT, unencrypted, unauthenticated)
- *Test Scenario 2 (Medium Security):* port 8080 (MQTT, WebSockets, unencrypted, unauthenticated)
- *Test Scenario 3 (High Security):* port 8091 (MQTT, WebSockets, encrypted, authenticated).

Ports were selected to validate our DevSecOps model over extremes, ranging from totally unprotected with unencrypted services and no authentication to authenticated and encrypted services.

4. Results

In the proposed DevSecOps model (Figure 3) the risks and vulnerabilities are found in seven stages. The reports of the scans will be sent to the user at every stage for analysis and the tools used here will also propose certain mitigation methods which can be incorporated to fix those security defects at the same stage. Thus, it aims at achieving security shift left. The vulnerabilities found in these activities are also tested for false positives and the defects are logged. The standardized alerting mechanism will send alerts and reports to the user after the security testing at each stage is completed and the dashboard will display the risk summary. The tables in this section represent the vulnerability count and the risk posture across all three ports. Results show that as the ports provide an increased level of security services there is a sharp decrease in the risk posture.

Likelihood	Level of Impact				
	Critical	High	Medium	Low	Informational
Critical	Critical	High	Medium	Low	Informational
High	Critical	High	Medium	Low	Informational
Medium	High	Medium	Medium	Low	Informational
Low	Medium	Low	Low	Low	Informational
Informational	Low	Low	Informational	Informational	Informational

Figure 4: NIST 800-30 Risk Matrix

The risk posture is calculated using NIST 800-30 revision1 standard which defines risk as: *the likelihood of an attacker exploiting that vulnerability and the impact caused by it in case it is exploited by a hacker* (see matrix in Figure 4). Each tool outputs the total number of security issues detected, in the form of the number of issues per risk category (e.g., x low, y medium, z high). The output from the tools contributes to the security risk rating by taking the number of issues detected in the highest risk level reported. When nothing is reporting from a tool, the low category is incremented by 1. The aggregate risk is determined by summing each category in the contribution security risk rating to identify which category has the highest number of reported issues.

Table1: Security risk posture of the product when MQTT on 1883 is used.

Security Posture			
DevSecOps Activity	Security Issues	Contribution	Aggregate Security Risk
Threat Modelling	5 High, 3 Medium, 4 Low	5 High	HIGH
Static Analysis	76 High, 1,477 Medium, 1,199 Low	76 High	
Vulnerability Scan	0 High, 3 Medium, 0 Low	3 Medium	
3rd Party	1 High, 0 Medium, 0 Low	1 High	
Dynamic Analysis	0 High, 2 Medium, 10 Low	2 Medium	
Penetration Testing	2 Critical, 0 High, 0 Medium, 0 Low	2 Critical	

Table 1 depicts the security posture after the DevSecOps process which was run on test.mosquitto.org -p 1883 (MQTT, unencrypted, unauthenticated). The overall risk posture is High due to issues identified during SDL activities. There are currently 83 High, 1486 Medium and 1214 Low open security issues which are present in the production. This is an expected result given that the configuration of the MQTT service had very low security in Test Scenario 1.

Table2: Security risk posture of the product when MQTT on 8080 is used

Security Posture			
DevSecOps Activity	Security Issues	Contribution	Aggregate Security Risk
Threat Modelling	Medium: 2	Medium: 2	Medium
Static Analysis	--	--	
Vulnerability Scan	High: 3, Medium: 1	High: 3	
3rd Party	High: 1, Medium: 2, Low: 1	High: 1	
Dynamic Analysis	Medium: 2	Medium: 2	
Penetration Testing	--	--	

Table 2 depicts the security posture after the DevSecOps process which was run on test.mosquitto.org -p 8080 (WebSockets, unencrypted, unauthenticated). The overall risk posture is Medium due to issues identified during SDL activities. There are currently 4 High, 7 Medium and 1 Low open security issues which are present in the production. This is an expected result given the configuration used in Test Scenario 2.

Table3: Security risk posture of the product when MQTT on 8091 is used

Security Posture			
DevSecOps Activity	Security Issues	Contribution	Aggregate Security Risk
Threat Modelling	--	--	Low
Static Analysis	--	--	
Vulnerability Scan	--	--	
3rd Party	Medium 1, Low: 3	Medium 1	
Dynamic Analysis	--	--	
Penetration Testing	0 High, 2 Medium, 0 Low	2 Medium	

Table 3 depicts the security posture after the DevSecOps process which was run on test.mosquitto.org -p 8091 (MQTT over WebSockets, encrypted, authenticated). The overall risk posture is Low due to issues identified

during SDL activities. There are currently 3 Medium and 3 Low open security issues which are present in the production, with 4 of the 6 phases not identifying any security issues. This is expected due to the higher security of the configuration used in Test Scenario 3.

The results show that all three test scenarios defined and assessed produced expected results given the specific security levels of each configuration. This demonstrated that our proposed MQTT DevSecOps model is valid and can be used to detect security issues at each stage of the development lifecycle, which can then be remediated before deployment.

5. Conclusion

With the increasing IoT devices and weak IoT communication protocols, cybercrime is costing more than ever. Among the several popular protocols for IoT communication MQTT stands in the top for several reasons such as its power to be scalable across millions of devices, lightweight, its efficiency, the reliability it offers and the speed of communication being extremely fast. However, the major setbacks include data privacy issues when the data is in transit, authentication, authorization and being controlled by botnet. This potentially exposes huge customer data in the open internet and attracts hackers due to the risks and vulnerabilities it includes. Fixing any such security defect which is in the open, especially in IoT products become highly expensive and cause huge loss if exploited. Therefore, a strong system is required wherein vulnerabilities can be found in the initial stages of the product cycle itself.

DevSecOps is one such methodology which involves security integration with enhanced automation and tooling at every stage of the product development lifecycle. It aims at finding vulnerabilities from the initial stages itself and helps in rigorous patching. There are multiple stages involved in this method such as threat modelling, secure code review SAST, DAST and penetration testing along with industry standard tools and frameworks which helps in achieving the job. These pipelines can be run along with the development pipelines on the need basis to spot the risks. It involves early threat prediction, detection, and alerting mechanism. Thus, it helps in achieving standard maturity for the IoT devices. We proposed a new model for DevSecOps in MQTT IoT systems and validated this on three different configurations using the open-source Mosquitto broker. Our results show that our model provides an effective pipeline for risk identification in these systems. Security issues were identified during SDL activities for test.mosquitto.org -p 1883, -p 8080 and -p 8091. SDL helps developers build more secure software and address security compliance requirements while reducing development cost. The product underwent various phases of the SDL and security issues were identified. The overall risk posture is High on ports 1883, Medium on 8080, and Low on 8091.

Future work could include a more performance aware pipeline development. That is, the introduction of any security tool or assessment incurs a delay overhead, which could breach the requirements of certain application releases. We propose to create a more sophisticated selection algorithm to create the pipeline by taking as input, the risk tolerance, release constraints, overhead of tools, etc. The ultimate goal is to create a tailored pipeline that respects the release constraints and minimizes the security/agility trade-off in DevSecOps.

References

- Bahaa, A., Abdelaziz, A., Sayed, A., Elfangary, L. and Fahmy, H., (2021). Monitoring Real Time Security Attacks for IoT Systems Using DevSecOps: A Systematic Literature Review. *Information*, 12(4), p.154.
- Casteur, G., Aubaret, A., Blondeau, B., Clouet, V., Quemat, A., Pical, V. and Zitouni, R., (2020), June. Fuzzing attacks for vulnerability discovery within MQTT protocol. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 420-425). IEEE.
- Chen, F., Huo, Y., Zhu, J. and Fan, D., (2020), November. A Review on the Study on MQTT Security Challenge. In 2020 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 128-133). IEEE.
- Ebert, C., Gallardo, G., Hernantes, J. and Serrano, N., (2016). DevOps. *Ieee Software*, 33(3), pp.94-100.
- Fox, M.R., (2020). IT Governance in a DevOps World. *IT Professional*, 22(5), pp.54-61.
- Hassija, V., Chamola, V., Saxena, V., Jain, D., Goyal, P. and Sikdar, B., (2019). A survey on IoT security: application areas, security threats, and solution architectures. *IEEE Access*, 7, pp.82721-82743.
- Harsha, M.S., Bhavani, B.M. and Kundhavai, K.R., (2018), September. Analysis of vulnerabilities in MQTT security using Shodan API and implementation of its countermeasures via authentication and ACLs. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2244-2250). IEEE.
- Hintaw, A.J., Manickam, S., Karuppayah, S. and Aboalmaaly, M.F., (2019). A Brief Review on MQTT's Security Issues within the Internet of Things (IoT). *J. Commun.*, 14(6), pp.463-469.

- Kenyon, H.S., (2021). DevSecOps Puts Security at the Heart of Program Development. *Defense AR Journal*, 28(2), pp.241-241.
- Kumar, R. and Goyal, R. (2021). When Security Meets Velocity: Modeling Continuous Security for Cloud Applications using DevSecOps. In *Innovative Data Communication Technologies and Application* (pp. 415-432). Springer, Singapore.
- Meneghello, F., Calore, M., Zucchetto, D., Polese, M. and Zanella, A., (2019). IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices. *IEEE Internet of Things Journal*, 6(5), pp.8182-8201.
- McAfee (2020). <https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf>
- Mishra, B. and Kertesz, A., (2020). The use of MQTT in M2M and IoT systems: A survey. *IEEE Access*, 8, pp.201071-201086.
- Myrbakken, H. and Colomo-Palacios, R., (2017), October. DevSecOps: a multivocal literature review. In *International Conference on Software Process Improvement and Capability Determination* (pp. 17-29). Springer, Cham.
- Okano, M.T., (2017), September. IoT and industry 4.0: the industrial new revolution. In *International Conference on Management and Information Systems* (Vol. 25, p. 26).
- Palo Alto (2020). <https://unit42.paloaltonetworks.com/iot-threat-report-2020/>
- Perera, P., Bandara, M. and Perera, I., (2016), September. Evaluating the impact of DevOps practice in Sri Lankan software development organizations. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 281-287). IEEE.
- Rajasekharaiah, C., (2021). Securing Microservices on Cloud. In *Cloud-Based Microservices* (pp. 179-202). Apress, Berkeley, CA.
- Sharma, C. and Gondhi, N.K., (2018), February. Communication protocol stack for constrained IoT systems. In *2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)* (pp. 1-6). IEEE.
- Syaiful, S., Rahardjo, B., and Hanindhito B. (2017) "Attack scenarios and security analysis of MQTT communication protocol in IoT system." In *4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 1-6. IEEE, 2017.
- Wolf, A., Simopoulos, D., D'Avino, L. and Schwaiger, P., (2021). The PASTA threat model implementation in the IoT development life cycle. *INFORMATIK* 2020.
- Zhu, L., Bass, L. and Champlin-Scharff, G., (2016). DevOps and its practices. *IEEE Software*, 33(3), pp.32-34.

Exploring Ontologies for Mitigation Selection of Industrial Control System Vulnerabilities

Thomas Heverin¹, Michael Cordano², Andy Zeyher¹, Matthew Lashner¹ and Sanjana Suresh¹

¹Drexel University, Philadelphia, USA

²Gryphon Technologies, Philadelphia, USA

th424@drexel.edu

Michael.cordano@gryphontechnologies.com

az458@drexel.edu

mil28@drexel.edu

ss5264@drexel.edu

Abstract: Mitigating vulnerabilities in industrial control systems (ICSs) represents a highly complex task. ICSs may contain an abundance of device types, all with unique software and hardware components. Upon discovering vulnerabilities on ICS devices, cyber defenders must determine which mitigations to implement, and which mitigations can apply across multiple vulnerabilities. Cyber defenders need techniques to optimize mitigation selection. This exploratory research paper shows how ontologies, also known as linked-data models, can potentially be used to model ICS devices, vulnerabilities, and mitigations, as well as to identify mitigations that can remediate or mitigate multiple vulnerabilities. Ontologies can be used to reduce the complexity of a cyber defender's role by allowing for insights to be drawn, especially in the ICS domain. Data are modelled from the Common Platform Enumeration (CPE), the National Vulnerability Database (NVD), standardized list of controls from the National Institute of Standards and Technology (NIST), and ICS Cyber Emergency Response Team (CERT) advisories. Semantic queries provide the techniques for mitigation prioritization. A case study is described for a selected programmable logic controller (PLC), its known vulnerabilities from the NVD, and recommended mitigations from ICS CERT. Overall, this research shows how ontologies can be used to link together existing data sources, to run queries over the linked data, and to allow for new insights to be drawn for mitigation selection.

Keywords: ontology, industrial control systems, vulnerability management

1. Introduction

Industrial control systems (ICSs) support the real-time control of large-scale operations and industrial processes in entities such as electricity distribution systems, water treatment facilities, nuclear power plants, manufacturing plants, smart cities, and military weapons systems. The National Institute of Standards and Technology (NIST) (2018, p. 98) states that ICSs consist of “combinations of control components (e.g., electrical, mechanical, hydraulic, pneumatic) that act together to achieve an industrial objective (e.g., manufacturing, transportation of matter or energy.”

The U.S. Cybersecurity and Infrastructure Security Agency (CISA) (2020) considers many types of ICSs to be “...so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, or any combination thereof.” As ICSs are vital to society, they have often been targeted by cyber attackers. One of the most well-known ICS cyber-attacks was Stuxnet. Stuxnet was the first weaponized cyber-attack against an ICS. The worm targeted Iranian nuclear plant centrifuges in 2010, causing substantial damage to the Iranian nuclear program (Knapp and Langhill, 2015). More recently, in the U.S there have been cyber-attacks on water treatment facilities, intrusions into dams, ransomware in meat processing plants, and breaches in car manufacturing plants. Cyber defenders of ICSs play a critical role. Cyber defenders must identify the assets (hardware and software) found in ICSs, determine vulnerabilities found on those assets, and select mitigations to protect the assets. Completing these steps can be quite time consuming and challenging. The goal of this research paper is to reduce the complexity of protecting ICSs.

2. Background

2.1 Complexity of ICS Vulnerability Management

ICSs are complex systems to defend for various reasons. Bristow (2020) stated that maintaining the inventory of all the assets found in an ICS represents a time-consuming and costly effort. With potentially hundreds of unique ICS devices on a network, along with traditional hardware and software assets, it can be a daunting task to keep

track of every asset on the network (Arnaert, Bertrand & Boudaoud, 2016). Veitch et al (2013) stated that ICSs lack built-in security features, lack mitigation technologies, and use protocols with many vulnerabilities.

Furthermore, with many devices included in an ICS, the attack surface of an ICS network can be considerably larger than a general company's network. The larger attack surface of an ICS network can make it easier for attackers to find openings to infiltrate the ICS. Additionally, ICS devices typically hold unique software and hardware which can be hard to maintain and upgrade. This means that in many cases, ICS devices are often unpatched leaving vulnerabilities exposed for years (Bristow, 2020).

In vulnerability management, cyber defenders must synthesize data from multiple sources to detect, respond and recover from attacks (D'Amico et al, 2005; Syed, 2020). To find and organize vulnerabilities, cyber defenders use specialized searches on the Internet; however, these searches can result in thousands of results that "...can be complex to sort, understand, and analyze." (Arnaert, Bertrand & Boudaoud, 2016 p. 299). With all the data that must be synthesized, cyber defenders face a cognitive overload which negatively impacts their work (D'Amico et al, 2005; Tyworth et al, 2012; Yen et al, 2010). Furthermore, cyber defenders work in high-stakes domains that require quick reaction time to mitigate vulnerabilities before attackers develop exploits (Joh and Malaiya, 2010). There are many demands on cyber defenders in protecting ICS networks, especially since cyber defenders must sift through many data sources. Ontologies can provide one way to help lessen these data-intensive demands. Ontologies provide the foundation for the discovery of new relationships within a domain, and across domains, especially when various data sources are linked together (World Wide Web Consortium, 2015).

2.2 Ontologies for Decision Making in Vulnerability Management

Ontologies have been identified as one way to integrate vulnerability data from multiple sources. Ontologies provide formalized, explicit knowledge representations of domains, including objects in those domains and relationships between those objects (Guariono, 1998; Syed, 2020). Ontologies can model relationships between data, integrate data, and allow for reasoning across the data (Wan, Guo & Camargo, 2010; World Wide Web Consortium, 2015). According to Syed (2020), although real-world vulnerability management involves integrating data from multiple vulnerability data sources, there is limited research in this area.

Wang and Guo (2009) developed an ontology for vulnerability management by modelling key entities and relationships such as vulnerabilities, attacks on the vulnerabilities, and countermeasures. It used the National Vulnerability Database (NVD) as its primary source. Fenz and Ekelhart (2009) created a general ontology of information security, with a small part focused on vulnerabilities. The ontology was modelled across various sources including NIST sources and others. Ontologies have been used to model vulnerabilities and requirements in the cybersecurity domain (Elahi, Yu, and Zannone, 2009). Vorobiev and Bekmamedova (2010) created several ontologies, one of which focuses on vulnerabilities of assets. They linked together cybersecurity policies with vulnerability mitigations.

Salini and Shenbagam (2015) included vulnerabilities in their ontology model to develop techniques for attack prediction. They used three models in their ontology: a threat model, a vulnerability model, and an attack model. Arnaert, Bertrand, and Boudaoud (2016) developed an ontology to model vulnerabilities of ICS objects to reduce the number of aggregated results, increase the relevance of results, and be usable for non-cybersecurity experts. They found that across various cybersecurity data sources, there lacks correlation between vulnerabilities and Computer Emergency Response Team (CERT) data elements. However, they only focused on one type of vulnerability (default credentials) in their research.

Syed (2020) developed a vulnerability management ontology to better inform users about impending vulnerabilities and countermeasures. Data sources included official cybersecurity data sources, along with social media intelligence. Although there have been multiple research papers authored about using ontologies for vulnerability management, Syed (2020) stated that critical items are missing from ontology-vulnerability models: the lack of context in vulnerability modelling (including CVSS scores) and a lack of linking to countermeasures to use against vulnerabilities. Our research aims to help fill these gaps in research on vulnerability management ontologies.

3. ICS Mitigation Ontology

3.1 Ontology Domain Concepts

In ontology development, it is critical for domain experts to identify domain concepts and relationships among the concepts (Sugumaran and Storey, 2002). The authors of this paper have over a decade of experience in the ICS cybersecurity domain, especially in the U.S. Navy ICS domain. In the vulnerability-management domain, main concepts include assets, vulnerabilities on those assets, and mitigations that can be used on those vulnerabilities (Syed, 2020). We used this common knowledge in vulnerability-management to explicitly link together ICS product, ICS vulnerability, and ICS mitigation data sources.

ICSs consist of assets including computers, human machine interfaces (HMIs), sensors, and programmable logic controllers (PLCs). PLCs represent critical ICS assets and play a central role in ICSs. ICS technicians configure PLCs to manage specialized inputs and outputs from devices such as actuators, valves, motors, engines, fans, pumps, and sensors (such as for geographic position, temperature, liquid levels etc). PLCs also consist of logic programs that automate the control and monitoring of industrial processes (Ghaleb et al, 2018). In the naval warfare domain, PLCs can be used to control many ship systems including steering, navigation, engineering, missile launch, radar, and gun systems. Given that PLCs play critical roles in ICSs, PLCs represent the focus of this study.

Vulnerabilities can be found on all types of assets. ICS assets, although critical assets, may have multiple vulnerabilities. According to NIST (2021), a vulnerability is “a weakness in the computational logic (e.g., code) found in software and hardware components that, when exploited, results in a negative impact to confidentiality, integrity, or availability.” Many vulnerabilities have been found on PLCs across various vendors (such as Siemens and Allen-Bradley Rockwell Automation).

Mitigations are protections that cyber defenders can use to remediate, or mitigate the risk of, identified vulnerabilities. Example mitigations include using a virtual private network (VPN), updating software to a newer version, restricting port access, minimizing network exposure, and more.

3.2 Ontology Data Sources

We explicitly model these three concepts (assets, vulnerabilities, and mitigations), as well as other concepts, in a Mitigations Ontology. To best represent these concepts, we use real-world data sources including sources that help standardize or provide asset names, vulnerabilities, and mitigations.

Since many different companies may call the same hardware or software asset by different names, the Common Platform Enumeration (CPE) specification provides a standardized way to name all types of assets including ICS assets (MITRE, 2013). As an example, the CPE standardized format for a S7-300 PLC is this: `cpe:2.3:h:siemens:simatic_s7-300`. By using CPE as a naming method for ICS assets, we can link ICS assets to various other data sources that also rely on the CPE nomenclature.

The National Vulnerability Database (NVD) collects and stores publicly known vulnerabilities. The NVD is one of the largest publicly-available vulnerability databases (Blinokwski and Piotrowski, 2020). NVD vulnerabilities are stored as Common Vulnerability and Exposure (CVEs) entries. CVE entries include a description of the vulnerability or exposure, as well as the type of vulnerability. These entries also include the product vendor and affected firmware or software versions in CPE format. CVE entries contain metadata in terms of the confidentiality, integrity, and availability-impact metrics as well as exploitability metrics. These metrics provide values according to the Common Vulnerability Scoring System (CVSS). The NVD utilizes the CVSS standard to rate vulnerabilities based on their potential impact to the affected system or software (Forum of Incident Response and Security Teams, 2021).

An example CVE found on the S7-300 PLC is CVE-2016-9158. For CVE-2016-9158, remote attackers can cause the PLC to go into defect mode, requiring a cold start of the PLC. The CVSS severity is score 7.5 (out of 10), the impact subscore is 3.6 and the exploitability subscore is 3.9

Although the NVD provides explicit data and metrics/scores about vulnerabilities, it does not directly provide mitigations for those vulnerabilities. It may provide links to external resources, such as vendor-specific bulletins, for vulnerabilities. In the ICS domain, cyber defenders must research ICS CERT advisories to find mitigations. ICS CERT advisories provide information including executive summaries, affected products, vulnerability overview,

links to CVEs, and mitigations (Cybersecurity and Infrastructure Security Agency, 2021). The mitigations in ICS CERT advisories are provided in full-text sentences or paragraphs which poses a challenge for ICS cyber defenders as they must spend time reading long paragraphs to determine which mitigations to use. NIST 800.53 (revision 5.1) provides a standardized list of controls (mitigations) one can use to secure information systems. However, ICS CERT mitigations that are described in full-text are not mapped to NIST 800.53 controls. Our ontology lays the foundation to expand on this lack of mapping.

3.3 Ontology Structure

Ontologies consist of three main types of entities: classes, object properties, and data properties. We used Protégé, a common ontology development tool, for our modelling. Classes represent the main concepts to be modelled, and signify categories of objects. Figure 1 shows the main classes in our Mitigations Ontology. For example, the “Vulnerability” class consists of CVEs from the NVD like CVE-2016-9158 while the “ICSDevice” class consists of device names like the S7-300 PLC.

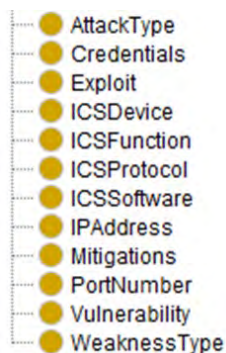


Figure 1: Classes in the Mitigations Ontology

The Mitigations class contains specific mitigations. As mentioned previously, ICS CERT advisories provide recommended mitigations in full-sentence and paragraph form. We manually developed a vocabulary of mitigations from analysing multiple ICS CERT advisories on PLCs. Then we mapped the sentences from ICS CERT advisories that described mitigations to our vocabulary of mitigations. Furthermore, we mapped our vocabulary to NIST 800.53 (revision 5.1) that provides a standardized list of controls. Two researchers independently developed vocabularies. Then those vocabularies were compared and finally synthesized into one final ICS Mitigation Vocabulary. The current ICS Mitigation Vocabulary is shown in Figure 2. The list in Figure 2 forms the objects of the Mitigations class.

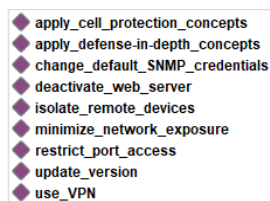


Figure 2: Objects Found in the Mitigations Class.

Object properties are used to formalize the links between classes and then to specify links between named objects within those classes. As an example, the object property *canMitigate* is the name of the object property that links together the classes “Mitigations” and “Vulnerability.” One can read this link as: “Mitigation *canMitigate* Vulnerability.” We can then “fill out” or specify this object property (*canMitigate*) at the object level. The following is an example of this:

- “deactivate_web_server” (a mitigation) *canMitigate* CVE-2016-9158 (a vulnerability).

When object properties are specified, they are considered to be triples. Triples consist of a subject, predicate, and object. When triples are formed, they are classified as being in the Resource Description Framework (RDF) format. RDF provides a formalized way to model with ontologies. Figure 3 shows examples of the object property *canMitigate* specified, linking specific mitigations with specific CVEs.

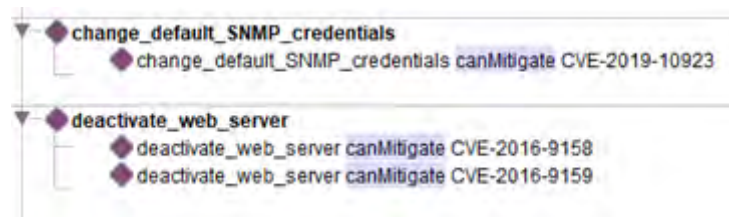


Figure 3: Examples of the CanMitigate Object Property in the Mitigations Ontology

With object properties we can create a chain of links. The below is an example showing a specific device linked with a specific vulnerability that is then linked to a specific mitigation:

- S7-300 PLC *hasVulnerability* CVE-2016-9158 and deactivate_web_server *canMitigate* this vulnerability.

Data properties provide metadata about objects in an ontology. They provide descriptive details about objects. As an example, Figure 4 shows key data properties for vulnerabilities that are drawn from the CVSS. For example, in regard to CVE-2016-9158, the CVSS base score is 7.5 (out of 10), the ImpactSubScore is 3.6 and the Exploitability Subscore is 3.9.

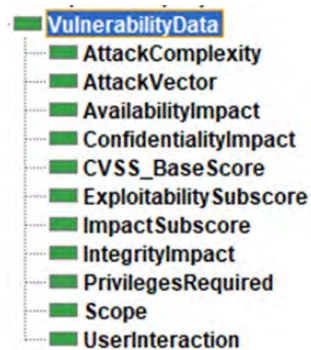


Figure 4: Vulnerability Data Properties Found in the Mitigations Ontology

We can also use data properties to provide names of objects in unique syntax such as CPE syntax. For example, we use a data property called “CPE” associated with ICS Devices. For the S7-300 this CPE data property is: cpe:2.3:h:siemens:simatic_s7-300. With Mitigation Ontology classes, object properties, and data properties specified, one can query over the ontology to draw insights and make decisions by using queries.

4. Ontology Queries and Results

4.1 SPARQL for Ontology Queries

SPARQL, a recursive acronym for SPARQL Protocol and RDF Query Language, is one common query language used to query RDF data (triples) in ontologies (World Wide Web Consortium, 2008). SPARQL holds many similarities with Structured Query Language (SQL). One advantage of using SPARQL is that, while being able to still query data out of relational databases like SQL, it is designed to combine vastly different data from several sources. In terms of syntax, SPARQL uses similar commands found in SQL, including SELECT, FROM, WHERE, UNION, GROUP BY, HAVING, and more. Furthermore, the results of SPARQL queries can be used with visualizations to allow for the development of insights. Protégé, provides OntoGraph, a visualization tool, to visualize relationships within ontologies. Examples of these are shown below.

4.2 SPARQL Queries for Vulnerability Management

In our queries, we first wanted to search across our Mitigations Ontology to identify all vulnerabilities associated with S7-300 PLC. The syntax is this SPARQL is provided below:

```
SELECT ?Asset ?Vulnerability
WHERE { Asset ics:hasVulnerability ?Vulnerability . FILTER (?Asset = ics:S7-300 ) .}
ORDERBY (?Vulnerability)
```

The above query produces 17 vulnerabilities (CVEs) that are associated with S7-300 PLCs. The results are visualized in Figure 5 via OntoGraph in Protégé. Each object in the Figure 5 can be examined by clicking on it to see additional object properties as well as data properties associated with that object.

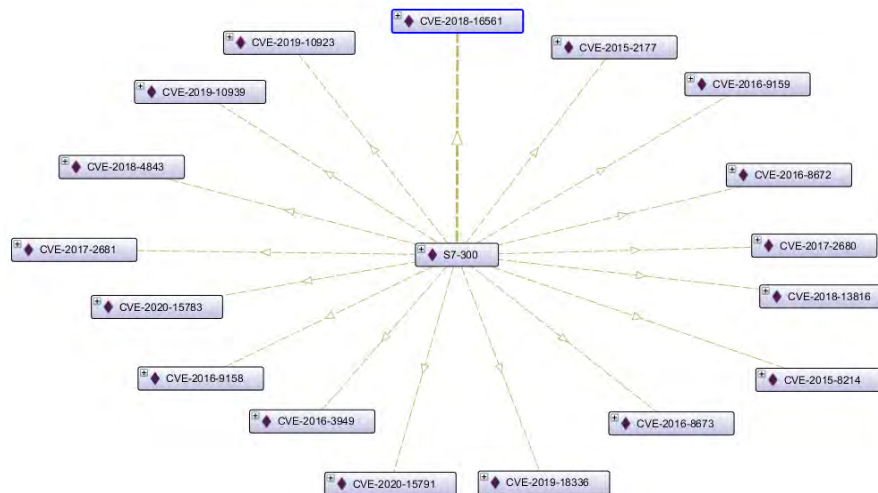


Figure 5: Visual of SPARQL Query Results that Show Vulnerabilities for S7-300 PLCs

When given a list of vulnerabilities, cyber defenders need prioritize which mitigations to use and determine, if possible, which mitigations could be used across multiple CVEs. Figure 6 shows how mitigations are mapped to S7-300 CVEs based off object property specifications (“canMitigate”) within the Mitigations Ontology.

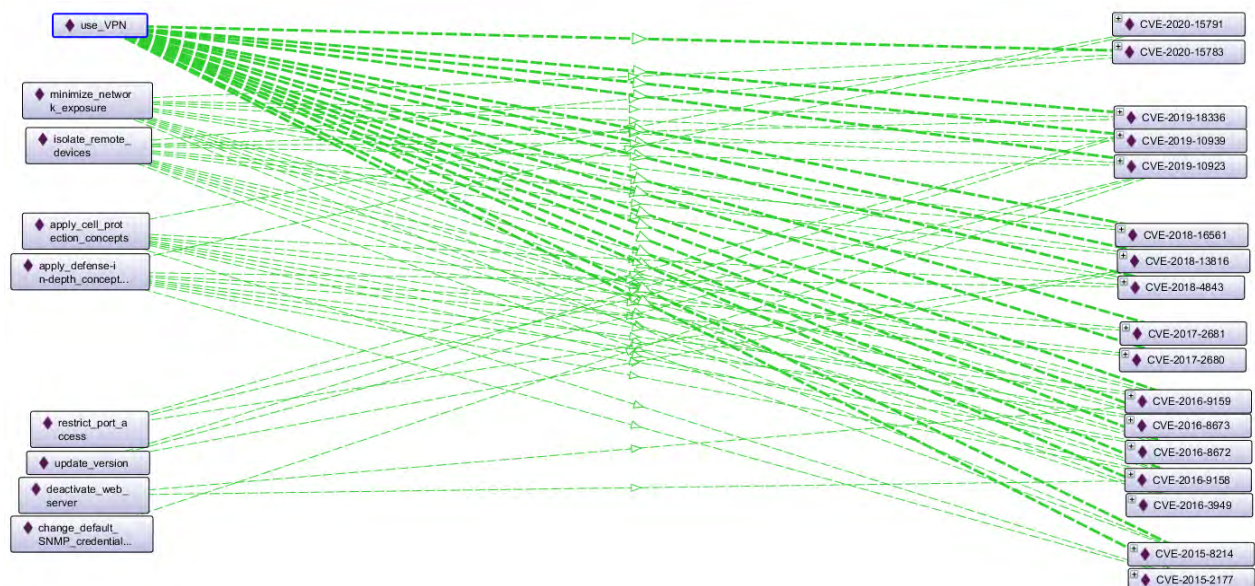


Figure 6: Visualization of Mitigations Mapped to S7-300 PLC CVEs

In Figure 6, on the left side, the mitigations are listed; on the right side, the associated CVEs are listed. The lines linking mitigations with the CVEs signify a specification of the object property “canMitigate.” For example the line between “use_VPN” and CVE-2020-15783 can be read as “use_VPN *canMitigate* CVE-2020-15783.” Various insights can be drawn from Figure 6. For example, one can see that the following mitigations map to a considerable number of CVEs: use_VPN, minimize_network_exposure, isolate_remote_devices, apply_cell_protection_concepts, and apply_defense_in_depth_concepts. The remaining mitigations only map to either one, two or three CVEs: restrict_port_access, update_version, deactivate_web_server, and change_default_SNMP_credential. Therefore, based on Figure 6, ICS cyber defenders can reason that using a mitigation from the top left will mitigate a considerable number of CVEs for the S7-300 PLC. This will help optimize the mitigation selection process.

Although ontology visualizations can allow for the drawing of insights, query results can provide empirical results to support the insights that are drawn. The next SPARQL query we developed, focused on figuring out which mitigations could be used across multiple S7-300 PLCs as well as the associated NIST 800.53 standardized controls. Also, each NIST 800.53 control falls within a control family or a general category of controls. Seeing the

control families can help cyber defenders determine what types of mitigation steps need to be taken. The syntax of this SPARQL query includes the following and the results are shown in Table 1:

```
SELECT ?Mitigation (COUNT(?Mitigation) as ?MitigationCount)
WHERE { ?Mitigation ics:canMitigate ?Vulnerability.
?Mitigation ics:mapsToNISTControl ?Control.
?Control ics:mapsToControlFamily ?ControlFamily .}
GROUPBY ?Mitigation ORDERBY DESC(?MitigationCount)
```

Table 1: SPARQL Query Results of Mitigations Across Vulnerabilities

Mitigations for S7-300 PLCs	Number of S7-300 PLC CVEs Mitigated	NIST 800.53 Rev. 5.1 Controls	NIST SP 800-53 Rev 5.1 – Control Families
Use VPN	16	AC-17 (Remote Access)	Access Control
Minimize network exposure	13	SC-7(2) (Boundary Protection – Access Points)	Boundary Protection
Isolate remote devices	13	SC-7(21)(Boundary Protection – Isolation of System Components)	Boundary Protection
Apply cell protection concepts	9	SC-1 (System Policy and Procedures)	System and Communications Protection
Apply defense in depth_concepts	9	PL-8(1) (Security and Privacy Architectures – Defense in Depth)	Planning
Restrict port access	3	AC-3(5) (Access Enforcement – Security Relevant Information)	Access Control
Update version	3	SI-7 (Software, Firmware, and Information Integrity)	System and Information Integrity
Deactivate web server	2	SC-7(19) (Boundary Protection – Block Communication)	System and Communication Protection
Change default SNMP credentials	1	IA-5(5)(Authenticator Management – Change Authenticators)	Identification and Authentication

The SPARQL results in Table 1 show that using VPN can mitigate 16 out of 17 vulnerabilities (94% of the vulnerabilities) associated with the S7-300 PLC. The visual in Figure 5, allows for user interaction. For example, upon clicking on CVE-2020-15971 (the one CVE that “use_VPN” does not mitigate), two mitigations are available: “apply_cell_protection_concepts” and “apply_defense_in_depth_concepts.” Selecting one of these mitigations along with “use_VPN” will mitigate 17 out of 17 vulnerabilities for the S7-300 PLC.

Additional insights can be drawn from Table 1. For example, we can see that, overall, “Access Control” and “Boundary Protection” mitigations can mitigate a considerable amount of S7-300 PLC vulnerabilities. Based on the real-world ICS cybersecurity experience of multiple authors of this paper, these two types of the controls are heavily used in the military ICS mitigations domain especially on in-service Navy ICSs. In other words, these types of mitigations can be used on operational ICSs. From Table 1, we also see that planning and architecture mitigations, “System Policy and Procedures” and “Planning”, can mitigate 9 out of 16 CVEs. The specific mitigations that fall under these two general categories include: “apply_cell_protection_concepts” and “apply_defense_in_depth_concepts.” Based on the authors' ICS cybersecurity experience, policies and planning also play a key role in ICS cybersecurity especially when planning the design of ICSs. However, these controls may not be used immediately in a live ICS environment. Therefore, an ICS defender working in a live environment may select other controls to immediately address S7-300 PLC vulnerabilities.

Vulnerability management, in terms of mitigations selection, is complex due to sources not directly mapped to each other. More work is needed in this exploratory research to find ways to help reduce the complexity of mitigation selection especially when examining multiple ICS devices.

5. Conclusion and Future Work

Our exploratory research shows that ontology modelling can be used to link asset, vulnerability and mitigation sources as well as shows that SPARQL queries can be run to aid in vulnerability management. SPARQL queries focused on vulnerabilities and their mitigations demonstrate that a list can be generated for specific mitigations that can be used across the most vulnerabilities. These results can aid ICS cyber defenders in their decision making, since ICS defenders work in high-stakes environments that require optimal use of resources and time. This work can be expanded in the future with more automation of data ingestion into the ontology. For example, in the current research project, ontology modelling was completed manually. Future research will focus on automating information extraction from the NVD to pull in more ICS vulnerabilities along with vulnerability metadata. Future research will also focus on using natural language processing (NLP) to read ICS CERT advisory sentences and paragraphs in order to map those sentences and paragraphs to our ICS Mitigations Vocabulary as well as to NIST 800.53 (revision 5.1) controls. We also will use testing and evaluation techniques including a task evaluation to assess the correctness of results and evaluation of the structure of the ontology by domain experts (Brank, Grobelnik, and Mladenić, 2005).

References

- Arnaert, M, Bertrand, Y, & Boudaoud, K 2016, 'Modeling vulnerable Internet of Things on SHODAN and CENSYS: An ontology for cyber security', *Tenth International Conference on Emerging Security Information, Systems and Technologies*, pp. 299-302.
- Blinowski, GJ & Piotrowski, P 2020, 'CVE based classification of vulnerable IoT systems', *International Conference on Dependability and Complex Systems, 2020*, pp. 82-93.
- Brank, J, Grobelnik, M & Mladenić D 2005, 'A survey of ontology evaluation techniques', *Conference on Data Mining and Data Warehouses, Ljubljana, Slovenia, 2005*.
- Bristow, M 2020, *ICS asset identification: it's more than just security*. SANS Institute.
- Cybersecurity and Infrastructure Security Agency 2020, *Critical infrastructure sectors*, viewed 12 October 2021, <<https://www.cisa.gov/critical-infrastructure-sectors>>.
- Cybersecurity and Infrastructure Security Agency, 2021, *ICS-CERT advisories*, viewed 12 October 2021, <<https://us-cert.cisa.gov/ics/advisories>>.
- D'Amico, A, Whitley, K, Tesone, D, O'Brien, B & Roth, E 2005, 'Achieving cyber defense situational awareness: a cognitive task analysis of information assurance analysts', *The Human Factors and Ergonomics Society Annual Meeting, 2005*, pp. 229-233.
- Elahi, G, Yu, E & Zannone, N 2009, 'A modeling ontology for integrating vulnerabilities into security requirements conceptual foundations', *International Conference on Conceptual Modeling, Gramado, Brazil 2009*, pp. 99-114.
- Fenz, S & Ekelhart A 2009, 'Formalizing information security knowledge', *The 4th International Symposium on Information, Computer, and Communications Security, Sydney, NSW, Australia, 2009*, pp. 183-194.
- Forum of Incident Response and Security Teams, 2021, *Common Vulnerability Scoring System Version 3.1: Specification Document*, viewed 12 October 2021, <<https://www.first.org/cvss/specification-document>>.
- Ghaleb, A, Zhioua, A & Almulhem, A 2018, 'On PLC network security', *International Journal of Critical Infrastructure Protection*, vol. 22, pp. 62-69.
- Guarino, N 1998, 'Formal ontology in information systems', *The International Conference on Formal Ontology in Information Systems, Trento, Italy, 1998*, pp. 3-15.
- Joh, H & Malaiya Y 2010, 'A framework for software security risk evaluation using the vulnerability lifecycle and CVSS metrics', *International Workshop on Risk and Trust in Extended Enterprises, San Jose, CA, 2010*, pp. 430-434.
- Knapp, ED, & Langill, JT 2015, *Industrial network security: securing critical infrastructure networks for smart grid, SCADA, and other industrial control systems*, Syngress, Waltham, MA.
- MITRE 2013, *Common Platform Enumeration*, viewed 16 October 2021, <<https://cpe.mitre.org/specification/>>.
- National Institute of Standards and Technology 2004, *Standards for security categorization of federal information and information systems*.
- National Institute of Standards and Technology 2018, *Risk Management Framework for information systems and organizations*. Revision 2.
- National Institute of Standards and Technology 2020, *SP 800-53 Release Search*, viewed 15 November 2021, <<https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/release-search#!/>>.
- National Institute of Standards and Technology 2021, *Vulnerabilities*, viewed 12 October 2021, <<https://nvd.nist.gov/vuln>>.
- Salini, P & Shenbagam J 2015, 'Prediction and classification of web application attacks using vulnerability ontology' *International Journal of Computer Applications*, vol. 116, no. 21, pp. 42-47.

- Sugumaran, V & Storey, V 2002, 'Ontologies for conceptual modeling: their creation, use, and management', *Data Knowledge Engineering*, vol. 42, no. 3, pp. 251–271.
- Syed, R 2020, 'Cybersecurity vulnerability management: A conceptual ontology and cyber intelligence alert system', *Information & Management*, vol 57, no. 6.
- Tyworth, M, Giacobe, NA, Mancuso, V & Dancy, C 2012, 'The distributed nature of cyber situation awareness.' *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2012.
- Veitch, CK, Henry, JM, Richardson, BT, & Hart, DH 2013, *Microgrid cyber security reference architecture* (No. SAND2013-5472). Sandia National Lab, Albuquerque, NM.
- Vorobiev, A & Bekmamedova N 2010, 'An ontology-driven approach applied to information security', *Journal of Research and Practice in Information Technology*, vol. 42, no. 1, pp. 61–76.
- Wang, JA & Guo, M 2009, 'Security data mining in an ontology for vulnerability management', *The International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, Washington, DC, 2009*, pp. 597–603.
- Wang, JA, Guo, MM & Camargo J 2010, 'An ontological approach to computer system security', *Information Security Journal: A Global Perspective*, vol. 19, no. 2, pp. 61–73.
- World Wide Web Consortium 2008, *SPARQL query language for RDF*, viewed 12 October 2021, <<https://www.w3.org/TR/rdf-sparql-query/>>.
- World Wide Web Consortium 2015, *Ontology*, viewed 12 October 2021, <<https://www.w3.org/standards/semanticweb/ontology>>.
- Yen, J, McNeese, M, Mullen, T, Hall, D, Fan, X & Liu, P 2010, 'RPD-based hypothesis reasoning for cyber situation awareness', in S Jajodia (ed), *Cyber Situational Awareness*, Springer, Boston, MA, pp. 39-49.

Future Needs of the Cybersecurity Workforce

Connie Justice¹, Char Sample^{1,2,3,4}, Sin Ming Loo⁴, Alex Ball⁴ and Clay Hampton¹

¹Purdue School of Engineering and Technology, IUPUI, Indianapolis, USA

²MTSI, Alexandria, USA

³SABSA Institute, UK

⁴Boise State University, Boise, USA

cjustice@iupui.edu

char.sample@mtsi-vai.com

Abstract: Expected growth of the job market for cyber security professionals in both the US and the UK remains strong for the foreseeable future. While there are many roles to be found in cyber security, that vary from penetration tester to chief information security officer (CISO). One job of particular interest is security architect. The rise in Zero Trust Architecture (ZTA) implementations, especially in the cloud environment, promises an increase in the demand for these security professionals. A security architect requires a set of knowledge, skills, and abilities covering the responsibility for integrating the various security components to successfully support an organization's goals. In order to achieve the goal of seamless integrated security, the architect must combine technical skills with business, and interpersonal skills. Many of these same skills are required of the CISO, suggesting that the role of security architect may be a professional stepping-stone to the role of CISO. We expected degreed programs to offer courses in security architecture. Accredited university cyber security programs in the United Kingdom (UK) and the United States of America (USA) were examined for course offerings in security architecture. Results found the majority of programs did not offer a course in security architecture. Considering the role of the universities in preparing C-suite executives, the absence of cyber security architecture offerings is both troubling and surprising.

Keywords: cyber security, security architect, CISO, education, workforce, Zero Trust Architecture (ZTA)

1. Introduction

Cyber Security education is facing an inflection point. Currently, there are 464,420 unfilled cybersecurity jobs (Cyberseek heatmaps) and the nature of the growth of these jobs suggests additional skills that are not being addressed in degreed cybersecurity programs. Cybersecurity is possibly in the strongest position to continue growing through proposed changes in course curricula, to accommodate changes in employers' needs. The research introduced in this manuscript compares curricula of UK and US universities to identify the percentage of universities with course offerings in cybersecurity architecture. A cyber security architect is a senior-level professional responsible for designing complex cyber security solutions synthesizing technology, management, and interpersonal skills within an organization (Cyber Degrees, n.d., Sherwood Applied Business Security Architecture, n.d.). Thus, the dearth of security architecture course offerings was not initially noticed. However, the growth of the cloud for site hosting along with the growing acceptance of Zero Trust Architecture (ZTA) implementation could exacerbate the skills gap.

Cyber Security programs continue development across universities creating their own academic silos in response to growing workforce demands for cyber security professionals (Cyberseek heatmaps, State of Cybersecurity, 2021, Cybersecurity Jobs Report, 2021). Strong industry growth justifies this increased pattern in cyber security programs. Current industry trends show millions of job openings by the year 2020/2021 (Cyberseek heatmaps). In ISACA's State of Cybersecurity Part 1. 61% of employers stated that they were significantly or somewhat understaffed (State of Cybersecurity, 2021). However, with the rise of artificial intelligence and machine learning (AI/ML), the distribution of jobs will likely change even if the number does not.

University cyber security educational programs continue turning out specialists supporting the market demand, creating entry-level penetration testers, assessors, auditors, and reverse engineers (Peltsverger 2015). Cybersecurity departments are common in most university curricula and will likely remain in place for some time. The field of cybersecurity has plenty of job openings providing university programs the opportunity to train students and enabling the achievement of high employment metrics.

However, cybersecurity professional demands continually change along with the virtual landscape. These changes require cybersecurity professionals to continually update their skillsets. Changes in both technologies and service offerings can outpace curricula development and deployment in many university programs. The time required to move from course envisioning to course delivery varies per institution but fails to keep pace with

industry's changing landscape. What is needed are course offerings in cybersecurity that teach students transferable skills (i.e., critical thinking, systems thinking) that can be applied in any situation, while allowing students to remain current in technologies and services.

The 2018 New Approaches to Cybersecurity Education (NACE) workshop in New Orleans, LA emphasized the need for diversity of thought with critical thinking skills regarding cybersecurity curricula (NACE, 2018, Cybersecurity Jobs Report, 2017). The workshop's papers recognized the need for cyber security programs to expand beyond the computer science departments, computer technology, computer engineering, and information science departments, and to even think beyond traditional science, technology, engineering, and math (STEM) disciplines. Hence, the recognition of the importance of both critical and adaptive thinking suggests cyber security programs, particularly those that offer an interdisciplinary approach, may be better positioned than others to adequately deal with the changing landscape and position students to become cyber security architects (Ramirez 2017).

A growing faction of cybersecurity scholars and academics have observed the need to break down silos and are also calling for interdisciplinary approaches to solving cyber security problems (Peltsverger 2015) that position students well in the areas that a well-rounded cyber security architect requires. Disciplines such as law, psychology, sociology, business management, resilience, reliability, statistics, data science, international studies, and others are becoming increasingly intertwined with cyber security (Ibid, 2015). Several initiatives exist to provide cyber security curricular guidance such as The Department of Homeland Security (DHS) and the National Security Agency (NSA) who jointly sponsored the National Centers of Academic Excellence (CAE) program and The Joint Task Force on Cybersecurity Education (CSEC2017), and the NIST National Initiative for Cybersecurity Education (NICE). These programs focus on curricula as an interdisciplinary path of study (Joint Task Force on Cybersecurity Education 2017, Cybersecurity Jobs Report 2017, National Initiative for Cybersecurity Education (NICE)). The response of the various university cyber security programs has been mixed. This response is understandable since agreement to break down the academic silos requires less effort than the act of integrating the disciplines.

One potential measure of a program's readiness to engage in updating curricula to reflect the impending new reality is the offering of security architecture courses. A Security Architect is defined for the purposes of this paper as a top-tier cybersecurity professional designing, directing, and overseeing cybersecurity and network security for the entire business entity (Cybersecurity Jobs Report 2017). Traditional architects combine knowledge from various disciplines in order to design structurally sound buildings (Savold et al. 2017). Similarly, security architects use skills learned in other disciplines to create robust network security solutions supporting organizational goals. Creating strong defensive networks in support of a mission requires a mix of breadth and depth in the skill set of the network architect (Triolo 2014). This research analysis of university cyber security programs in the US and UK provides the findings of one measure of preparedness, the offering of security architecture classes, for future requirements.

2. Background

2.1 Cybersecurity Workforce

Study international (2021) not only showed continued growth for all of cybersecurity but specifically mentioned skills that require building in security across all layers, "from the ground up" (Ibid 2021). This differs from the old approach where professionals were simply told to 'think like a hacker' (Sternstein 2012). The problem with the 'think like a hacker' approach is that in many cases students learned to think like a hacker at the expense of thinking like a defender, a good cybersecurity professional needs to do both.

In addressing this need, The National Initiative for Cybersecurity Education (NICE) led by the National Institute of Standards and Technology (NIST) formed a partnership of government, academia, and the private sector (National Initiative for Cybersecurity Education (NICE)). The partnership's mission is to increase the number of cybersecurity professionals to secure our national infrastructure by building on existing cybersecurity successful programs, highlighting innovation, and collaborating with all stakeholders to ensure current and future cybersecurity challenges are met. NICE has four communities of interest, Apprenticeships in Cybersecurity, Cybersecurity Skills Competitions, K12 Cybersecurity Education, and NICE Framework Users. The purpose of the NICE communities of practice is to provide a means for private and public sectors to come together and drive

change advancing cybersecurity education, training, and workforce development (National Initiative for Cybersecurity Education (NICE)).

2.2 The NiCE Framework

The NICE Framework perceives competencies as identified by employer or educator and assesses the students' ability to meet the competency (e.g. do the work) figure 1. Consequently, competencies are grouped by task, knowledge, and skill (TKS) statements. Therefore, competencies are employer-driven, focus on the student, and are observable and measurable.

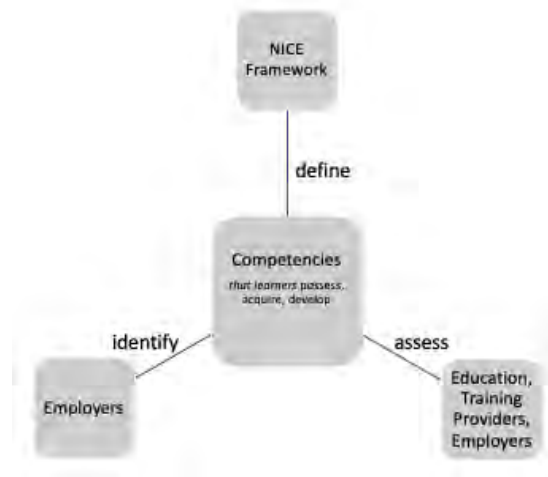


Figure 1: NICE Framework (NICE, n.d.)

The Association for Computing Machinery (ACM) assembled a Joint Task Force on Cybersecurity Education (CSEC2017 2017) to address the need for education to aid in workforce development. CSEC2017 is based on several curricular frameworks and is designed in part using the U.S. National Research Council Next Generation Science Standards (CSEC2017 2017). The CSEC2017 views cybersecurity education grounded in scientific, computing, and cybersecurity principles that are continuously refined using evidence-based practice (CSEC2017 2017). Figure 2 shows the CSEC thought model. The knowledge areas depict the critical knowledge over a broad range of computing disciplines, and it is the basic structure of the model. The knowledge areas represent all knowledge within the cybersecurity discipline. The CSEC 2017 links to the NICE workforce framework as cybersecurity educational needs are unique due to the balancing of breadth in overall cybersecurity knowledge, depth in a specific area, and aligning the educational needs with workforce demands (Raj et.al. 2018).

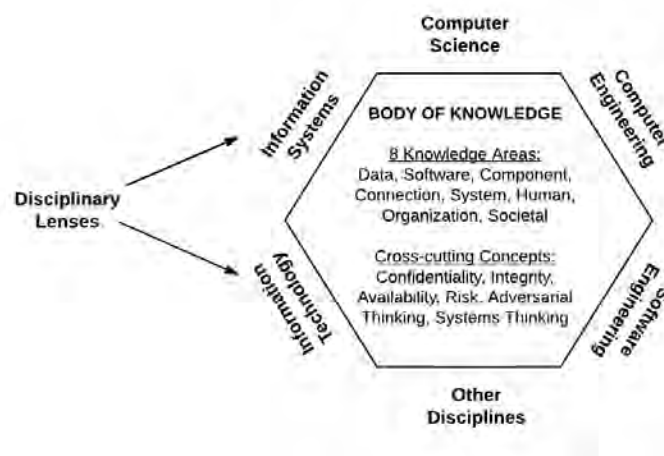


Figure 2: CSEC 2017 Thought Model (CSEC2017, 2017)

Another example of meeting the need for the cyber workforce is the National Security Agency (NSA), Department of Homeland Security's (DHS) partnership in the Centers of Academic Excellence (CAE). The CAE supports NICE in creating a pool of skilled workers able to protect the nation. Hiring skilled cybersecurity professionals to protect critical infrastructures, information systems, and networks is a top priority for

governments and businesses globally (Centers for Academic Excellence (CAE)). DHS provides tools, resources, and education aimed specifically at building the strength and sustainability of the cyber security workforce. DHS is working hard to retain current talent while filling the gaps by offering competitive bonuses and salaries in line with private businesses. Additionally, DHS provides a workforce tool for all companies to utilize identifying critical gaps in cyber security personnel (Centers for Academic Excellence (CAE)). The DHS Workforce development Lifecycle depicted in figure 2 is comprised of:

- Cybersecurity workforce identification and quantification
- Understanding workforce needs and skills gaps
- Hiring the right people for clearly defined roles
- Enhancing employee skills with training and professional development
- Creation of programs and experiences to retain top talent



Figure 3: DHS Workforce Development Lifecycle

More recently cloud migrations have resulted in a need for security architects. Organizational migration to the cloud is not new, Amazon, Microsoft, and Google have had offerings for years (Miller 2016; Harvey 2017; Foley 2018). Cybersecurity professionals were initially hired to identify risks in the cloud environment; however, more recently these same professionals have been asked to create solutions that enhance business workflows to allow for secure business processing. These solutions require the professional to go beyond thinking like a hacker, to thinking like a security architect, which requires an interdisciplinary skill set and mindset.

2.3 Academia's Response to Workforce Demand

One reason for the existence of academia silos can be explained through the research on the building of expertise that occurs by focusing on a specific discipline at the cost of exclusion of others. Studies are purposefully tightly restrained to allow the researcher to focus on a specific problem. Variables are limited, so results or findings can be generalized for applications where the same variables appear in different environments. Thus, cyber security, a discipline with underpinnings in computer science and mathematics, would naturally follow the same structural pattern. This ultimately leads to cyber security professionals who are conversant in technical speech but are less able to effectively comprehend or communicate with other disciplinary groups (i.e., management, marketing, etc.) found a typical work environment. In many cases, cybersecurity professionals are hackers who are quite skilled at breaking into systems and networks, but less skilled at building secure solutions (Bishop 2018).

Cyber security programs responded to the industry's demand for specific technical skillsets, and this approach showed initial successes resulting in the global growth of new programs. However, like nursing where professionals initially took care of the patient's immediate needs, programs evolved to include increasing numbers of courses and disciplines (psychology, chemistry, sociology, kinesiology, etc.) in order to better prepare nurses for their jobs. So too, cybersecurity curricula must evolve to include additional disciplines with the goal of improving students' overall performance for the future workplace, through understanding the various viewpoints necessary to architect strong security solutions.

Teaching students how to architect strong security solutions expands the training boundaries beyond technology to understand business processes, organizational and individual behaviors. The security architect must be able to put forth the various views of the solution system, that covers how each user understands their interaction

with the designed security solution while executing their role in the business process workflow (SABSA, DoDAF, TOGAF.). The interdisciplinary requirements associated with security architecture demands more than traditional technology knowledge from the cybersecurity professional. University programs are tasked with creating agile thinkers, yet many of the exercises focus more on tactics, tools, and procedures which are more consistent with training rather than reflective thought.

Security architecture work continues to change in support of growing security requirements that reflect the cloud-based architectures. The interdisciplinary nature of security architecture work is well-suited for course development and delivery that blend technical and behavioral disciplines. This mix of skills is compounded in the virtual environment which offers flexible capabilities, but these capabilities also introduce additional security challenges (Rose, Borchert et al., 2020). One way of creating security solutions in this cloud is the adoption of Zero Trust Architecture (ZTA) concepts. ZTA's assumption of a compromised hostile environment presents additional challenges to designing and integrating new secure architectures (Ibid). Thus, students and course developers have an opportunity to use and refine interdisciplinary skills (law, finance, psychology, sociology, data science, computer science, information science, etc.). However, students need coursework that teaches the inner workings of technologies, business processes, and other disciplines in order to design ZTA compliant solutions.

This failure to adequately support other disciplines further isolates cyber security professionals and may limit students to becoming industry commodities. Commodities are quickly acquired and discarded, making for career-limited growth options. AI enabled tools promise to exacerbate the commodity problem. Robots powered by AI to perform hacking are a cheaper option than humans. To ensure cybersecurity students are educated and trained to meet the needs of the cybersecurity workforce it is crucial to structure cybersecurity programs for future interdisciplinary growth.

3. Problem Statement

The interdisciplinary nature of cybersecurity increasingly suggests the need to restructure cyber security programs away from the silo approach and into the interdisciplinary approach. The general problem facing educational institutions, and students is accredited programs may not adequately prepare students for cybersecurity workforce challenges where diverse skill sets are becoming increasingly important. Specifically, the universities are focusing on technical rather than the holistic education of the cybersecurity learner when the workforce has a growing need for the holistic cybersecurity professional-the cyber security architect (Cyber Degrees). Universities need to develop a holistic cyber security architect program to meet the future cybersecurity demands in the job force.

The interdisciplinary nature of security architecture provides an attractive means to fuse together the traditional technology programs with other disciplines. In order to determine the best technical solution to a problem, the problem must first be well understood. Security architects are tasked with bringing together the various stakeholders to gain a greater understanding of their views and risk tolerances, *before* making technical recommendations (Sherwood Applied Business Security Architecture, n.d., DoDAF Architecture Framework Version 2.02, n.d.). This experience in practical application of solutions learned through various courses in both the undergraduate and graduate programs exists in some universities but not all programs. This led the researchers to ask the question, how many programs offer security architecture courses?

4. Research Method

This study with the resultant analysis is exploratory in nature. The goal is to determine the rate of programs offering security architecture courses. Thus, this study is a preliminary data collection effort over 2 two countries (US & UK) that seeks to answer a simple question, what percentage of university-based cybersecurity programs offer a course in security architecture? The researchers hope to use this initial study as a launching point for additional studies in education course curricula. The study is limited to university programs that are accredited by their respective country's accreditation agencies. Certificate programs, junior colleges, and community colleges were exempted from this study.

One reason why this study was limited to university programs is the knowledge requirement of a security architect (CSEC2017, Peltsverger 2015). As previously mentioned, security architecture requires knowledge of both security technology, information technology in general, and business processes required to support the

organization's mission (CSEC2017, Peltsverger 2015); therefore, the professional engaged in this line of work will possess professional experience. The experience associated with security architecture should not serve as a deterrent to the various cyber security programs since all programs surveyed offered courses in security policy and management of security professionals, both areas with strong experience requirements. Data collection relied on querying the individual institutionally owned websites discovered through the countries' accrediting agencies. Furthermore, training courses while offering good technology experiences lack the much-needed context that provides the experiential component that students desperately need.

5. Results

All data presented was based on data obtained from the UK and US accredited cybersecurity programs with data extracted in October 2021. What was discovered was that many universities did not call the courses cyber security architecture, nor did they call the course security architecture. Few universities offered cyber security architecture or security architecture courses. If we noticed any of those courses in the data, we noted the name. The overall rate for US accredited universities was 215 and the overall rate of UK universities was 75. The sample set for the US was N=215 and UK N=75. 13 universities in the US offered a security architecture course and 202 universities had no security architecture course. Conversely, four universities in the UK offered security architecture courses and 69 did not offer a security architecture course.

2021

Country	Architectural courses offered				Course not offered	Unable to discern	Total
	General	Enterprise	Computer	Security			
UK	0	1	0	4	69	1	75
UK (%)	0%	1.33%	0%	5.33%	92%	1.33%	
US	41	22	98	13	40	1	215
US%	19.10%	10.20%	45.60%	6.00%	18.60%	0.47%	

Other interesting discoveries were made through the course of this research. There were accredited programs in cybersecurity that did not possess a degree in cybersecurity or courses in security architecture. Some of the architecture courses that were not counted did not fit the definition of security architecture course such as skills learned in other disciplines to create robust network security solutions supporting organizational and a mix of breadth and depth in the skill set of the network architect (Triolo 2014). The courses were more hardware architecture oriented or software architecture oriented. There were courses such as systems architecture, network architecture, PC architecture, and software architecture. These courses did not have anything in common with the security architect course that is recommended by the authors. Additionally, some programs showed research only security focus and no focus on the interdisciplinary security required of the security architect (Ramirez 2017).

5.1 Proposed Solutions

There are several potential solutions to the cyber security architect problem and each one warrants discussion. The proposed solutions are not limited to those discussed here and are likely highly situational. In some cases, some institutions may find some programs unworkable, for this reason, these are suggestions, not requirements.

- Create a liaison position in the departments that interacts with other disciplines. This approach would entail hiring a liaison who reaches out to different departments and works to define the necessary courses to make cyber security a joint major with the available disciplines.
- Embed departments together to work on a common goal. An example of this approach occurs at Cardiff University in Wales where criminal justice, cyber security, data science, psychology, computer science exists in teams that work together in solving common research problems. Another example can be seen at Bournemouth University where the psychology department is co-located with the cyber security department where research into decision science in cyber security explores and answers questions in risk and trust related behaviors (www.bournemouth.ac.uk).
- Require cyber security to be a dual major or joint major at the undergraduate level. This would force cyber security students to understand how cyber supports other disciplines and communicate with personnel in a manner demonstrating an understanding of the discipline.
- Create distinct curriculum for cybersecurity majors including, but not limited to; cybersecurity risk assessment, creating policies, third party risk, and network security architecture.
- Create a cybersecurity department for all disciplines and require students to take the common core cybersecurity curriculum before taking curricula in specific disciplines. See figure 3. Additionally, create

common cybersecurity core curriculum before discipline specific curriculum and midway or end of discipline specific curriculum, see figure 4.

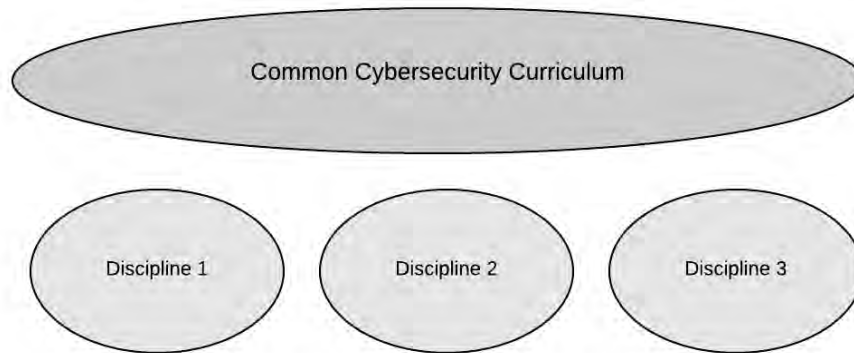


Figure 4: Common cybersecurity curriculum before specific discipline curriculum

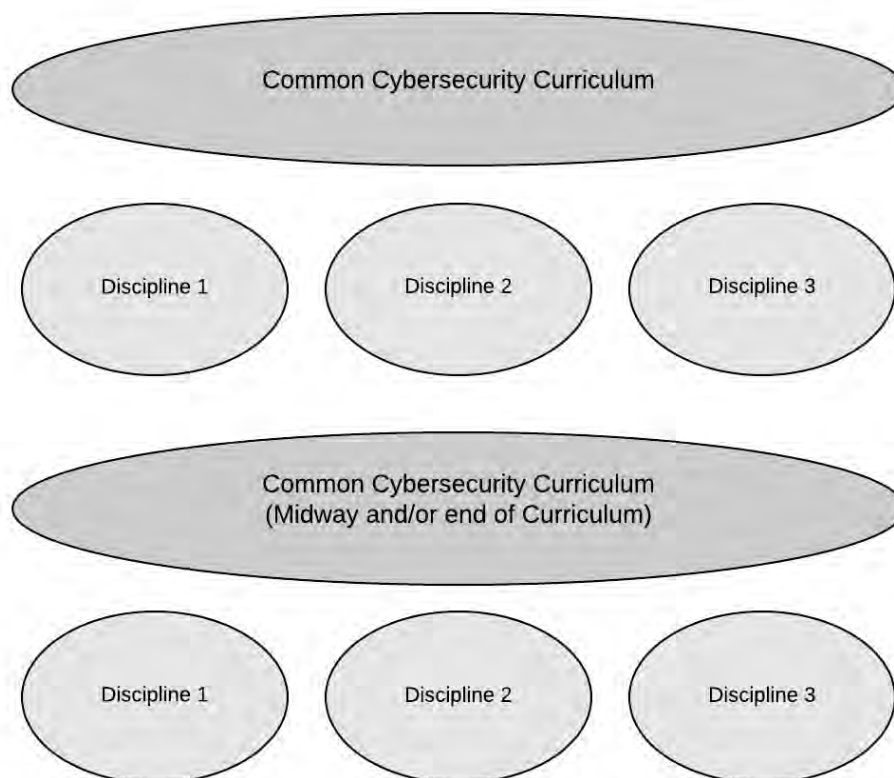


Figure 5: Common cybersecurity curriculum before and midway and/or end of specific discipline curriculum

5.2 Cybersecurity Bachelor's Degree (BS)

A new cybersecurity Bachelor of Science degree was created as a result of the research conducted in this manuscript. The degree was approved in the spring of 2021 and was available to students in the fall of 2021. The degree was based on figure 4 above with a common cybersecurity curriculum and disciplines/concentrations. The degree was initially built with courses that exist in the program, with the intent to build out concentrations in the coming semesters.

The plan of study is clustered as follows: primary cybersecurity courses, foundational cybersecurity courses, elective cybersecurity courses, and general undergraduate courses. The BS in cybersecurity is a 120 undergraduate credit requirement. The breakdown of the 120 credits is as follows:

Students must complete:

- 27 credits of Primary Cybersecurity courses (Concentration Courses)
- 36 credits Foundational cybersecurity courses
- 9 credits from the combined lists of primary cybersecurity, elective cybersecurity, and foundation cybersecurity courses
- 38 credits from general undergraduate courses and electives
- 10 credits of Leadership courses

Currently, there is only one concentration, the technical cybersecurity concentration. During the fall of 2021, two to three concentrations will be planned for the fall of 2022 semester. Ideas for concentrations to be considered are policy, education, psychology, ethics, communication, and linguistics. Along with the concentration, courses that will be added to the cybersecurity BS. are cyber security architecture and zero trust architecture.

5.3 Security Architecture Design

Security professionals are frequently reminded to “bake in” security, not “bolt it on” (Applegate, 2013). The IT industry still has many challenges in designing security from the start. Oftentimes, security is addressed after the fact and the system is heading towards production. Also, there ought to be better guidance for individuals creating requests for proposals (RFP), where the government sends out RFPs with security being addressed at later times in the design process. This security by design must be engineered to the environment, and the security solution supports processes from the very start. Typical cyber security architects in the field gain knowledge and experience from on-the-job experiences and this takes several years to attain. Designing security into the project requires other disciplinary knowledge outside of the traditional technical areas.

5.4 Robust Offensive and Defense Skills

Specialized roles such as penetration testers and reverse software engineers provide an entry point into an organization, but not professional growth opportunities. Triolo (2014) noted attackers need to be correct once and defenders need to be correct every time. A certain set of skills must bridge the gap between attacker skills and defender skills, these skills are often found in integrators who may have obtained the skills through experience or coursework.

5.5 Capture the Flag Training and Exercises

Many universities and colleges participate in capture the flag cyber challenges requiring participants to act as both attackers and defenders (Manson & Pike 2014; Murphy & Murphy 2013). These exercises are primarily focused on vulnerability exploitation, with prevention being covered as a reaction to attack signatures (Manson & Pike 2014; Murphy & Murphy 2013). In some cases, the cyber challenges require teams to build resilient solutions, but once again these solutions are designed to withstand known attacks in general. Creating and building defenses, in this arrangement, becomes an ad-hoc process that lacks rigor.

As an example, in the NACE workshop Bishop (2018) proposed creating “build the flag”, a new twist on capture the flag competitions as a way to broaden and strengthen the knowledge, skills, and abilities of participants. The idea of build the flag would enhance the cyber security architect curriculum by building entire systems from the ground up to meet specific requirements given. The focus is to design and build systems with security in mind bringing together all learning acquired (New Approaches to Cybersecurity Education (NACE) Workshop, 2018). The build the flag exercise that incorporates ZTA principles could serve as a competition or even a capstone project for students that would offer an alternative capable of appealing to non-traditional cyber students.

5.6 Management Skills

In Henry’s, “Mastering the cyber security skills crisis: realigning educational outcomes to industry requirements” paper, he discusses the Cyberspace Education Framework in which broad high-level educational objectives were narrowed to show key outcomes for specific cybersecurity activities (Henry, 2017). Such activities were then applied to cybersecurity knowledge, skills, and abilities (KSAs) in management. The important takeaway from this approach is that cybersecurity is multi-faceted, multi-dimensional, and purpose specific. The topics of study

included high-level and mid-level management as well as project management. According to Trilling 2018, cybersecurity professionals are usually excellent in technical skills, and then as they progress in their occupation, they are moved into managing without any training or education in the discipline. Businesses require skilled and prepared managers.

5.7 People Skills

The ability to communicate with people and articulate cybersecurity problems and issues; educate the workforce in cybersecurity dangers to protect company assets; and communicate and work with other employees and customers to improve the security posture of the company. Soft skills, specifically communication, was described as a critical skill that a cybersecurity professional needed to possess (Hartley et. al. 2017; Jones et. al. 2018; Yuan et. al. 2017). Jones et. al's paper asked the specific question concerning what soft skill is important to your job and communication was number one. Under communication was written communication, public speaking, presenting, and collaboration (Jones et.al 2018).

6. Recommended topics for the Cyber Security Architect

"Security architects design, build, and oversee the implementation of network security for an organization" (Cyber Degrees). The security architect is entrusted to create a solution reflecting a deep technical knowledge of security products, applications, and operating systems, and how to integrate those products in support of organizational goals. Solutions are complex and must work (Ibid). The mix of technical skills, management skills, and people skills are unique. Introducing this mix of skills in cyber security programs as a foundational course would provide a foundation for a wider path of experiences for students. This section introduces some concepts and/or ideas for content relevant for cyber security architects such as security architecture design; robust offensive and defensive skills; and capture the flag training and exercises. This is just a representation and is not inclusive of all that is available.

7. Conclusions

The changing nature of problems requiring interdisciplinary approaches to cyber problems will force change in educational institutions' programs. Higher education is in danger of falling behind in the race to educate future cybersecurity professionals (Knapp et.al. 2017). These changes will need to recognize the importance of other academic disciplines in creating the next generation of cyber security professionals. This paper identified and measured a gap in cyber security education and put forth suggestions to offer potential ways forward by educating students on cyber security architecture.

A potential side benefit to adding coursework in security architecture for students is that they are better prepared to speak in terms that potential employers understand while still maintaining their technical terms. These skills would, in turn, make students more valuable to employers. Institutions that succeed in this endeavor will also distinguish themselves with employers

8. Future Work

Future work would include extending the to all universities in the UK and in the US, not just the programs that are accredited by their respective country's accreditation agencies. Additionally, it would benefit the research if it were to include certificate programs, junior colleges, and community colleges as part of the study. Further future work on this topic will include defining curriculum across multiple disciplines for the cyber architect. Universities and colleges will need to develop programs from the undergrad to the doctorate level to meet the cyber architect degree requirements. Cyber architect work should be reviewed for the role across multiple industries including the DOD space.

Another item for consideration is reviewing the role of the system engineer and morphing this role into a cyber architect position. A review of current system engineer degrees and curriculum should be performed to determine the path to a cyber architect. System engineers typically move to a cyber position, so it would be beneficial to universities and colleges to review this path as well.

Another aspect that will create new AI related jobs in the cybersecurity field and for cyber architects is the concept of the Fourth Industrial Revolution where quantum computing, AI, precision medicine, 3-D printing, AI, autonomous vehicles, Nanotechnology, Neurotechnology; LIFI, high-speed bi-directional network/mobile communications using light, energy capture, storage, and transmission; geoengineering; blockchain;

development of advanced materials; additive manufacturing (3D Printing) and multidimensional printing; virtual and augmented realities; and space technologies will create new technical jobs (Schwab 2018). These new technological endeavors singular or combined will create new jobs and expand technological responsibilities for old ones. All of which will create new cybersecurity functions and applications.

References

- Become a security architect, Cyber Degrees, Retrieved from <https://www.cyberdegrees.org/jobs/security-architect/>
- Cybersecurity Jobs Report: 3.5 Million Openings In 2025. (n.d.). Retrieved November 14, 2021, from <https://cybersecurityventures.com/jobs/>
- Cybersecurity Skills Gap, 2016. Centers for Academic Excellence (CAE) <https://www.nsa.gov/resources/students-educators/centers-academic-excellence/>
- Cyberseek Heat Maps <https://www.cyberseek.org/heatmap.html>
- DoDAF Architecture Framework Version 2.02. <https://dodcio.defense.gov/library/dod-architecture-framework/>
- Hartley, R., Medlin, D., and Houlik, Z., 2017. Ethical Hacking: Educating Future Cybersecurity Professionals. In Proceedings of the EDSIG Conference ISSN (Vol. 2473, p. 3857).
- Harvey, C. 2017, May 25. "Google cloud platform: History features & pricing" website: <https://www.datamation.com/cloud/google-cloud-platform/>
- Henry, A., 2017. "Mastering the cyber security skills crisis: realigning educational outcomes to industry requirements" ACCS discussion, Australian Centre for Cyber Security, UNSW, paper no. 4, August 2017.
- Joint Task Force on Cybersecurity Education, Cybersecurity Curricula 2017 (CSEC2017). Available: <https://www.acm.org/binaries/content/assets/education/curricula-recommendations/csec2017.pdf>
- Jones, K.S., Namin, A.S. and Armstrong, M.E., 2018. The Core Cyber-Defense Knowledge, Skills, and Abilities That Cybersecurity Students Should Learn in School: Results from Interviews with Cybersecurity Professionals. ACM Transactions on Computing Education (TOCE), 18(3), p.11.
- Knapp, K. Maurer, C. and Plachkinova, M., 2017. "Maintaining a Cybersecurity Curriculum: Professional Certifications as Valuable Guidance." Journal of Information Systems Education 28(2): 101-113, 2017.
- Manson, D., and Pike, R., 2014. "The case for depth in cybersecurity education." ACM Inroads 5(1): 47-52, 2014.
- Murphy, D., and Murphy, R., 2013. "Teaching cybersecurity: Protecting the business environment," Proceedings of the 2013 on InfoSecCD '13: Information Security Curriculum Development Conference. Kennesaw GA, USA, ACM: 88-93, 2013.
- Miller, R., (2016, July 2). "How AWS came to be" website: <https://techcrunch.com/2016/07/02/andy-jassys-brief-history-of-the-genesis-of-aws/>
- National Initiative for Cybersecurity Education (NICE) <https://www.nist.gov/itl/applied-cybersecurity/nice/about>
- New Approaches to Cybersecurity Education (NACE) Workshop, 2018, <https://www.cerias.purdue.edu/site/nace/>
- National Initiative for Cybersecurity Careers and Studies (NICCS), Workforce Development, <https://niccs.us-cert.gov/workforce-development>
- Peltsverger, S., 2015. "A survey of university system of Georgia cyber security programs". Proceedings of the 2015 Information Security Curriculum, 2015.
- Raj, R.K., Blair, J.R., Sobieski, E. and Parrish, A., 2018, December. Enhancing Cybersecurity Content in Undergraduate Information Systems Programs: A Way Forward. In Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy (Vol. 1)
- Ramirez, R., Making cyber security interdisciplinary: recommendations for a novel curriculum and terminology harmonization, Massachusetts Institute of Technology, 2017, unpublished.
- Rose, S., Borchert, O., Mitchell, S., and Connelly, S., (2020 August). "SP 800-207 Zero Trust Architecture" website: <https://csrc.nist.gov/publications/detail/sp/800-207/final>
- Sherwood Applied Business Security Architecture (SABSA) 2021 website: <https://sabsa.org>
- Savold, R., Dagher, N., Frazier, P., and McCallam, D., "Architecting Cyber Defense: A Survey of the Leading Cyber Reference Architectures and Frameworks," In Cyber Security and Cloud Computing (CSCloud), 2017 IEEE 4th International Conference on (pp. 127-138). IEEE, 2017.
- Schwab, K. 2018. Shaping the fourth industrial revolution. Geneva, Switzerland: World Economic Forum.
- State of Cybersecurity 2021, Part 1: Global Update on Workforce Efforts, Resources and Budgets. (2021).
- Sternstein, A., 2012, January 30. "Feds need to start thinking like hackers" website: <https://www.nextgov.com/cybersecurity/2012/01/feds-need-to-start-thinking-like-hackers/50540/>
- Study International Staff, 2021, May 6. "Cybersecurity: A career that will stay in-demand for decades to come" website: <https://www.studyinternational.com/news/cybersecurity-career-in-demand/>
- Trilling, R., 2018, September. Creating a New Academic Discipline: Cybersecurity Management Education. In Proceedings of the 19th Annual SIG Conference on Information Technology Education (pp. 78-83). International World Wide Web Conferences Steering Committee.
- Triolo, C., "Hackers only need to get it right once, we need to get it right every time," SC Media, 2014.
- United States Department of Defense https://dodcio.defense.gov/Library/DoD-Architecture-Framework/dodaf20_viewpoints/

Yuan, X., Yang, L., He, W., Ellis, J.T., Xu, J. and Waters, C.K., 2017, March. Enhancing Cybersecurity Education Using POGIL. In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (pp. 719-719). ACM.

Zero Trust and Advanced Persistent Threats: Who Will Win the War?

Bilge Karabacak and Todd Whittaker

Franklin University, OH, US

bilge.karabacak@franklin.edu

todd.whittaker@franklin.edu

Abstract: Advanced Persistent Threats (APTs) are state-sponsored actors who break into computer networks for political or industrial espionage. Because of the nature of cyberspace and ever-changing sophisticated attack techniques, it is challenging to prevent and detect APT attacks. 2020 United States Federal Government data breach once again showed how difficult to protect networks from targeted attacks. Among many other solutions and techniques, zero trust is a promising security architecture that might effectively prevent the intrusion attempts of APT actors. In the zero trust model, no process insider or outside the network is trusted by default. Zero trust is also called perimeterless security to indicate that it changes the focus from network devices to assets. All processes are required to verify themselves to access the resources. In this paper, we focused on APT prevention. We sought an answer to the question: "could the 2020 United States Federal Government data breach have been prevented if the attacked networks used zero trust architecture?" To answer this question, we used MITRE's ATT&CK® framework to extract how the APT29 threat group techniques could be mitigated to prevent initial access to federal networks. Secondly, we listed basic constructs of the zero trust model using NIST Special Publication 800-207 and several other academic and industry resources. Finally, we analyzed how zero trust can prevent malicious APT activities. We found that zero trust has a strong potential of preventing APT attacks or mitigating them significantly. We also suggested that vulnerability scanning, application developer guidance, and training should not be neglected in zero trust implementations as they are not explicitly or strongly mentioned in NIST SP 800-207 and are among the mostly referred controls in academic and industry publications.

Keywords: Zero Trust, Zero Trust Architecture, ZTA, NIST SP 800-207, MITRE ATT&CK, Advanced Persistent Threats, APT, APT29, SolarWinds Orion Breach

1. Introduction

On December 13, 2020, Cybersecurity and Infrastructure Security Agency (CISA), the federal body responsible for managing the cybersecurity of critical sectors in the US, released an emergency directive (CISA, 2020). The directive was the announcement of the worst cyber-espionage incident ever suffered by the US because of the long duration and the sensitivity and high profile of the targets (Wikipedia, 2021a). If printed, the amount of data stolen would form a stack far taller than the Washington Monument (Bajak, 2020). It was probably one of the most organized cyber-espionage campaigns and the biggest cyber-raid against the US in years. The breach has been named in various names, including SolarWinds Orion Breach, SolarWinds Orion Software Supply Chain Breach, or 2020 United States Federal Government Data Breach.

Many federal organizations in the US were affected by this security incident, such as eleven Departments, including State, Defense, Treasury, Energy, Homeland Security, and Justice. US federal government was not the only victim. There are other Governments, Fortune 500 companies, International Organizations. For example, the British government, NATO, European parliament, and Microsoft were among the victims (Harding and Sabbagh, 2020). SolarWinds announced that it sent an advisory to about 33,000 of its customers around the globe who might have been affected, and roughly 18,000 customers had installed the compromised product (Cimpanu, 2020).

In this modern trojan horse story, cyber attackers turned SolarWinds Orion software into a trojan horse by injecting malicious code. To do that, they penetrated Solarwind's update server by obtaining the easy-to-guess password of the server and injected malware into Orion's update packages. SolarWinds was not the only exploited product; the attackers also exploited software or credentials from Microsoft and VMware.

Attackers started their malicious activities in September 2019. They spent roughly six months on reconnaissance and resource development before initiating the attack. During this time, they created their proof-of-concepts codes, developed a command-and-control system that would allow them to connect victim networks, send commands, and download data. They sneaked their code to the SolarWinds update server as early as March 2020; this is also considered the start day of penetrating federal networks in the US. It has been passed roughly nine months between the initial access of the attackers and the announcement of the breach by CISA. Eventually, the malicious activities span a timeline of fifteen months.

The cybersecurity advisory released by the United States holds "Russian Foreign Intelligence Service (SVR) actors" responsible for compromising SolarWinds Orion software update packages (NSA, CISA and FBI, 2021). The advisory specifically mentioned the names of three APT groups: APT29, Cozy Bear, and the Dukes.

When it comes to the information on the activities of these threat groups, the MITRE ATT&CK webpage helps security researchers and companies a lot. "MITRE ATT&CK is a knowledge base of cyber adversary behavior and taxonomy for adversarial actions across their lifecycle" (MITRE, 2021b). The tactics and techniques used by the APT29 group have been listed and analyzed on the MITRE ATT&CK webpage. Among the information listed for APT29, there are tools used by the group, procedures -specific implementations of techniques-, associated groups. ATT&CK framework shows Cozy Bear and The Dukes as associated groups with APT29. According to MITRE ATT&CK, APT29 uses more than 80 attack techniques spread into ten different tactics across the adversarial lifecycle. MITRE ATT&CK knowledge base is a community-supported and free-to-use threat intelligence database that can help detect malicious activities and attribute them to the threat groups.

Detection of cyber threats is essential in network defense; prevention is the first step, though. Cyber defenders, including the industry, non-profit organizations, government agencies, and academia, have been working on new services, software, models, tools, and architectures to prevent the attacks at their earliest phases. Zero trust network architecture is a coordinated cybersecurity and system management strategy that combines many existing cybersecurity solutions and paradigms with a different security mindset (NSA, 2021). Zero trust is not a siloed or isolated security application; instead, it is a model that changes the approaches to the trust concept in networks. In the zero trust model, the trust is removed from the network and systems; so that no process (user, computer, device) is trusted by default regardless of its network location and ownership (Rose *et al.*, 2020). Zero trust is also called perimeterless security to indicate the fact that it changes the focus from network devices to data and enterprise resources, and all processes are required to verify themselves to access the resources (Wikipedia, 2021b). Zero trust has been adapted by some enterprises, introduced by security vendors in their services and product portfolio, proposed by academia, and recently formalized by the US government by NIST Special Publication 800-207. Zero trust architecture can be very effective in preventing APT attacks and reducing the impact of sophisticated attacks considerably.

In this article, we reviewed the literature and MITRE ATT&CK database to answer the following questions:

1. Could the 2020 United States Federal Government Data Breach caused by SolarWinds Orion and other vendor vulnerabilities have been prevented if the attacked networks used zero trust architecture?
2. If not, what other controls would have been required to prevent the breach in federal networks?

The article's goal is to provide insights into the effectiveness of Zero Trust Architecture (ZTA) to prevent APT attacks. For that purpose, we reviewed four techniques and two sub-techniques used by the APT29 group to gain initial access to networks (MITRE, 2021a). We also reviewed NIST Special Publication 800-207 and academic resources to extract the main constructs of ZTA. Finally, we discussed the effectiveness of ZTA constructs in preventing the techniques and sub-techniques of the APT29 group and answered two research questions above.

The paper has four sections. The first section is the introduction. The second section is reserved for the literature review. The third section is the part of the article where we answer the research questions. The fourth section is dedicated to future studies and conclusions.

2. Literature Review

This section is grouped into three subsections. The first subsection is dedicated to the methods and tools to prevent APTs. The second subsection is reserved for the MITRE ATT&CK summary. The third subsection gives coverage to the literature about ZTA.

2.1 APT Prevention

As research questions imply, APT detection is not within the scope of this article. Therefore, the literature review on APT is limited to prevention.

It would not be wrong to say that most academic papers about APTs discuss detection methods instead of prevention methods. Vulnerabilities are inevitable. State-sponsored cyber actors are targeted groups and have been using sophisticated techniques. As an expected consequence of these facts, APT groups frequently compromise networks and systems (Karabacak and Tatar, 2014). As a matter of fact, successful APT attacks are

assumed to be successful no matter when. The incapability of prevention methods by nature might be the root reason behind the high proportion of APT detection papers (Tatar, Karabacak and Gheorghe, 2016). Cole also assumes the prevention failure in his book and says, "Prevention is ideal, but detection is a must" (Cole, 2013). Cole emphasizes the importance of both prevention and detection in fighting against APTs. According to Cole, companies usually forget about prevention and focus on detection.

On the other hand, the reverse is also true when looking at the long time spans of incidents; companies invest in prevention and show minimal effort on detection (Cole, 2013). The recent breach statistics confirm that millions of records have been stolen from organizations unnoticed for months, even years. Cole emphasizes the importance of augmentation of traditional prevention technologies with additional technologies. The author lists three traditional prevention technologies: firewall, Intrusion Prevention System (IPS), and Data Loss Prevention (DLP). For the augmentation of the firewall, he proposes adding internal firewalls from different vendors in addition to perimeter firewalls and using host-based firewalls. IPS augmentation steps are minimizing false positives and tuning to detect the indicators of APTs. DLPs can be augmented by tying with a Digital Rights Management solution. Finally, Cole indicates that these foundational technologies will not be sufficient to stop APTs.

Moon et al. propose a multi-layer defense system to prevent and detect APTs (Moon *et al.*, 2014). The defense system collects and analyzes log information from endpoints and processes the information with the help of eight components. The authors share two cases for the prevention: infection through USB and spear phishing. The proposed defense system might effectively prevent some specific cyberattacks; however, it cannot be applied to all circumstances. Mohamed et al. stress two crucial social aspects to prevent APT attacks: security awareness and information security policies (Mohamed, Jantan and Abiodun, 2018). They propose an enhancement to the MITRE ATT&CK Mitigations for a specific incident caused by malware infection. They propose automatic termination of the connection between the victim and attacker. Their proposed solution is not comprehensive and mature in mitigating a diverse set of APT tactics and techniques. Messaoud et al. provide a list of technologies to protect against APT actors. Those technologies are sandboxing, honeypots, SIEM, and user behavior analytics (Messaoud *et al.*, 2016). They map these technologies with different attack phases of APT actors in their lifecycle using a matrix. According to the proposed matrix, all these technologies are effective in preventing APT attacks; however, they don't share insights into the possible limitations of these four technologies in preventing APT attacks.

In addition to recommendations and solutions proposed by academia, security vendors continuously improve their product portfolio that help prevent APT attacks directly or indirectly. Firewall, antivirus, IPS, web application firewall, web & email protection, and sandboxing are among the most common products to prevent APT attacks (Hudson, 2014).

2.2 MITRE ATT&CK

MITRE ATT&CK is a publicly available threat intelligence database that groups the techniques used by cyber actors into 14 tactics sorted by the lifecycle of an attack (MITRE, 2021b). It is based on real-world observations. ATT&CK framework categorizes adversarial behavior into two broad classes; Enterprise and Mobile. There are customized ATT&CK matrices that fall into these two classes. When this article has been prepared, there are 11 matrices for enterprise networks and two matrices for mobile. There is also one matrix for industrial control systems. In addition to ATT&CK matrices that include all documented tactics and techniques for enterprise networks and mobile platforms, MITRE ATT&CK Navigator shows specific techniques used by a threat group. For this article, the authors reviewed six initial access techniques of the APT29 threat group after switching to the MITRE ATT&CK Navigator of the group. Tactics and techniques are the building stones of the framework. Techniques are listed under tactics; there are 188 Techniques, 379 Sub-techniques grouped under 14 tactics. "Techniques represent "how" an adversary achieves a tactical goal by performing an action. Tactics represent the "why" of an ATT&CK technique or sub-technique; they are the reason for performing an action. Sub-techniques are a more specific description of the adversarial behavior used to achieve a goal. They describe behavior at a lower level than a technique (MITRE, 2021b)." ATT&CK website shares other cyberthreat information such as threat groups and tools used by threat groups. The latest version of the ATT&CK (v10) provides detailed information about 129 threat groups and 637 pieces of software used in cyber-attacks.

There are many use cases that ATT&CK can be utilized. For example, it can be used

1. by defense teams to detect cyber threats
2. to model threats and perform comparable and structural threat intelligence
3. by red teams to emulate cyber threats
4. by security architects and engineers to assess the current security posture and redesign the security architecture.

As an example of the fourth use case, the MITRE organization shares the details of a behavioral-based threat model based on MITRE ATT&CK to perform validation of security products (Strom *et al.*, 2017). Several academic studies propose original security methods and models using the ATT&CK framework—the proposed methods and models by academia map to one of four use cases listed above. For example, Outkin *et al.* present a game-theoretic method that uses the MITRE ATTA&CK APT3 threat data to model attacker-defender interaction and enhance the defender strategies (Outkin *et al.*, 2021). Manocha *et al.* propose a security assessment rating framework using ATTA&CK (Manocha *et al.*, 2021). Pell *et al.* study a dynamic threat modeling for 5G networks using ATTA&CK (Pell *et al.*, 2021). Choi *et al.* show how ATT&CK can be used to generate random attack sequences against ICS datasets (Choi, Yun and Min, 2021). Georgiadou *et al.* evaluate organizational/individual security culture and security vulnerabilities together and map them to adversaries using ATT&CK to develop a cybersecurity culture framework (Georgiadou, Mouzakitis and Askounis, 2021). Xiong *et al.* propose a new threat modeling language for enterprise security based on the ATT&CK enterprise matrix (Xiong *et al.*, 2021).

The flexibility and open nature of ATT&CK attract the attention of enterprises and security vendors. Enterprises use the database to improve their security capabilities. Security product vendors actively use ATT&CK to enhance their product features and libraries and map their functionalities to the adversary techniques.

2.3 Zero Trust Architecture (ZTA)

“Zero trust is a cybersecurity paradigm focused on resource protection and the premise that trust is never granted implicitly but must be continually evaluated” (Rose *et al.*, 2020). In the zero trust model, no process insider or outside the network is trusted by default. Zero trust is also called perimeterless security to indicate that it changes the focus from networks, network devices, and perimeters to data and enterprise resources; all entities must verify themselves to access the resources. Some vendors develop zero-trust platforms, whereas some use zero-trust principles to improve their existing security solutions. Colortokens is one of the vendors that developed a zero-trust platform, and they explain how their platform helps stop APT threats (Nayak, 2021). A number of vendors have been preparing guidance documents to provide general information about ZTA and how their products adapt to the architecture (Chowdhury, 2019; Banach, 2021; Checkpoint, 2021; Colortokens, 2021; Splunk, 2021; VMware, 2022).

There are not so many academic studies that analyze the effectiveness and capabilities of ZTA in preventing APT attacks. Alevizos *et al.* claim that authenticated and authorized endpoints can be considered weak points in networks adopting ZTA so that APTs can perform malicious activities through these endpoints. They suggest using blockchain to verify the integrity of endpoints in a ZTA (Alevizos, Ta and Hashem Eiza, 2021). Horne and Nair claim that ZTA is a realistic security model to battle modern threats, including APTs (Horne and Nair, 2021). Although it is not an academic work, Van Driel provides an overview of how zero trust frustrates APT actors in his blog post (Van Driel, 2015).

Academic studies on ZTA have proposed solutions for specific cybersecurity and technology cases that do not directly touch APT prevention. Collier and Sarkis compare zero trust concepts and current supply chain concepts (Collier and Sarkis, 2021). They discuss the application of zero trust for more secure supply chains. Zaheer *et al.* propose eZTrust for microservices, a network-independent perimeterization approach using zero trust concepts (Zaheer *et al.*, 2019). One of the use cases for eZTrust is to deploy contingency policies that block the traffic to vulnerable applications. Tao *et al.* propose a method to perform security controls on big data (Tao, Lei and Ruxiang, 2018). They base their method on zero trust; the first phase of their three-phased approach is user context recognition based on zero trust. Chen *et al.* propose a security protection framework for power mobile Internet services using zero trust (Chen *et al.*, 2021). Mehraj and Banday propose a zero-trust strategy for the cloud environment (Mehraj and Banday, 2020). The strategy is in the conceptual design phase and is based on choosing trustworthy entities in the cloud environment. De Silva *et al.* propose a mathematical model for a zero-aware smart home system based on continuous authentication and context-aware access control; both are based on zero trust (da Silva, Macedo and dos Santos, 2021). Vanickis *et al.* propose a risk-based access control

framework for zero trust networks and define the policy languages -such as firewall access control policies- to support the framework (Vanickis *et al.*, 2018).

Trust itself is a vulnerability, and zero trust is a systemic approach to patch this vulnerability (Campbell, 2020). Zero trust might be very efficient in preventing today's sophisticated cyber attackers when implemented adequately by organizations. Academia collaboratively and continuously researches zero trust and proposes conceptual or technical solutions for various cases and systems, including cloud environments, smart homes, supply chains, big data, microservices, and power mobile internet services. Industry adapts itself very well to ZTA; they develop new services and products and share white papers with their partners and customers.

2.4 The Gap in the Literature

There are quite a few academic studies that bring the ZTA and APT concepts together. One of those studies proposes a ZTA and blockchain-based solution that focuses on preventing APTs (Alevizos, Ta and Hashem Eiza, 2021). To the best of our knowledge, there are no studies that research the effectiveness of ZTA in preventing APT attacks.

3. The Analysis of the Effectiveness of Zero Trust in Preventing APTs

In the next section, we analyze the effectiveness of ZTA in preventing APT actors. First, we use MITRE ATT&CK to extract and review the initial access techniques and sub-techniques of the APT29 group. Second, we use the literature and NIST SP 800-207 to extract the main building blocks of zero trust. Third, we analyze the effectiveness of ZTA in preventing the APT29 techniques.

We have several delimitations and limitations. First, we delimit our review to the 2020 United States Federal Government Breach. In this regard, we review the attack techniques of the APT29 group, as that group was one of the main actors of the government breach (NSA, CISA and FBI, 2021). The amount of information about APT actors is limited. MITRE ATT&CK provides the most comprehensive information about the techniques and tools used by APT actors. We used MITRE ATT&CK as the only source of information for APT29 techniques. APT29 group might be using more techniques than provided by the ATT&CK framework. In our reviews, we are limited by the information provided in the ATT&CK. We also assumed that no further tactics in the attack lifecycle would be successful without initial access to the networks.

This section has three sub-sections. In the first subsection, we share our review of four techniques and two sub-techniques of the APT29 group to gain "initial access" to networks. In the second subsection, we summarize the principles and building blocks of zero trust architecture. In the third subsection, we evaluate the effectiveness of zero trust in preventing APT techniques.

3.1 APT Analysis

We limited our scope to the initial access tactic of APT29 because our focus is the prevention at the very beginning of the threat lifecycle. If defenders succeed in thwarting APT actors at the initial access step, APTs will not be able to step up to the forthcoming tactics. Specifically, 11 tactics in ATT&CK Enterprise Matrix that come after the initial access are actions of the APT actors who completed the initial access. The two tactics before the initial access are reconnaissance and resource development. The preventive countermeasures associated with these tactics are within the scope of cyber deterrence and cyber diplomacy.

There are tools and methods, including zero trust, that can effectively prevent APT techniques associated with the tactics after the initial access. However, our focus is to analyze the effectiveness of zero trust in preventing APT activities without any network compromises.

Figure-1 shows the four techniques and two sub-techniques used by APT29 in initial access. Figure-1 is taken from the ATT&CK Navigator page of APT29. The shaded techniques are the ones used by APT29.

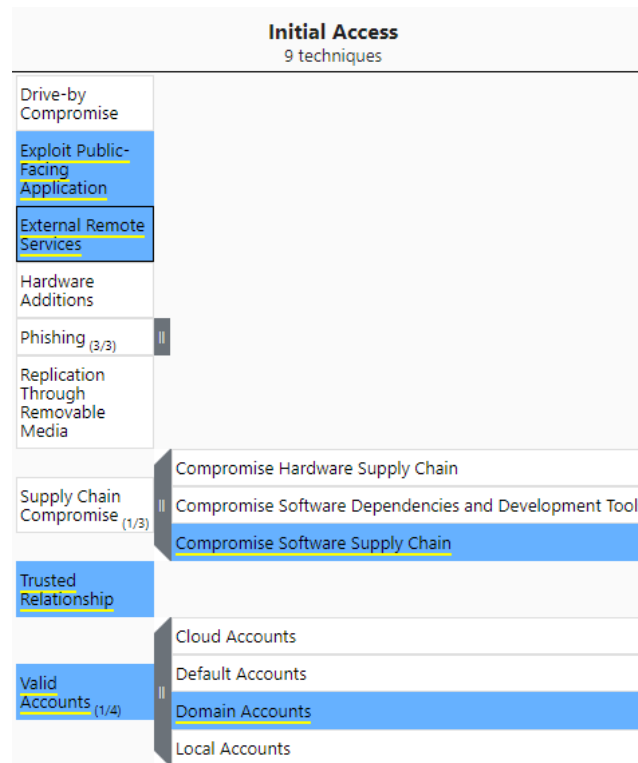


Figure 1: Initial access techniques and sub-techniques of APT29

Respective countermeasures for these techniques are listed in Table-1. The countermeasures are extracted from the ATT&CK webpage as well. The description of techniques and mitigations is not given in this article because of the space constraints.

Table 1: Mitigations for APT29 techniques and sub-techniques

Technique/Sub-technique	Mitigation
Exploit public-facing application	1. Application Isolation and Sandboxing (1) 2. Exploit Protection (2) 3. Network Segmentation (3) 4. Privileged Account Management (4) 5. Update Software (5) 6. Vulnerability Scanning (6)
External remote services	1. Disable or Remove Feature or Program (7) 2. Limit Access to Resource Over Network (8) 3. Multi-factor Authentication (9) 4. Network Segmentation (3)
Compromise software supply chain	1. Update Software (5) 2. Vulnerability Scanning (6)
Trusted relationship	1. Network Segmentation (3) 2. User Account Control (10)
Valid accounts	1. Application Developer Guidance (11) 2. Password Policies (12) 3. Privileged Account Management (4) 4. User Training (13)
Domain accounts	1. Multi-factor Authentication (9) 2. Privileged Account Management (4) 3. User Training (13)

After removing the duplicated mitigations from Table-1, we are left with 13 unique controls to compare with Zero Trust. Note that unique identifiers are put next to each control in the second column of Table-2.

3.2 Zero Trust Architecture

ZTA is not a straightforward solution; instead, it combines different kinds of technologies, methods, and procedures. It involves people, processes, and technology. Successful implementation of ZTA requires a total

commitment from leadership (Chowdhury, 2019; NSA, 2021). It should also be aligned with business objectives (Chowdhury, 2019). Every organization can adapt to ZTA differently depending on different organizational and technological factors. In this section, we extracted seventeen components of ZTA by reviewing various academic and industry resources (Modderkolk, 2018; Chowdhury, 2019; Rivas, 2019; Embrey, 2020; Rose *et al.*, 2020; Uttecht, 2020; Alevizos, Ta and Hashem Eiza, 2021; Banach, 2021; Buck *et al.*, 2021; Checkpoint, 2021; Colortokens, 2021; Garbis and Chapman, 2021; Sanders *et al.*, 2021):

1. Centralized authentication of users, devices, and applications (accompanied by SSO and MFA)
2. Encryption
3. Continuous monitoring
4. Network segmentation
5. Application segmentation
6. Network access control
7. System access control
8. Traffic filtering
9. Application execution control
10. Operational and forensic analysis (Includes vulnerability scanning (Uttecht, 2020))
11. Policy enforcement
12. Configuration management
13. Data discovery, tracking, and analysis (Asset inventory)
14. User account management
15. Least privilege principle
16. Guidance and training of application developers
17. Training and awareness

ZTA uses the existing technologies; the differentiation of ZTA is how it brings these technologies together to create a zero-trust mindset in organizations. Below is the list of fundamental tenets of zero-trust as shared in NIST SP 800-207. The above list of countermeasures might not help readers understand how ZTA is designed and deployed. The implementation context for the seventeen items above is explained thoroughly in NIST SP 800-207.

3.3 Comparative Analysis of ATT&CK Mitigations and ZTA Components

This section compares two numbered lists of countermeasures in Sections 3.1 and 3.2. The first list was the mitigations suggested by MITRE ATT&CK to prevent APT29 initial access attempts. The second list was the list of components of ZTA, composed after reviewing NIST SP 800-207 and several academic articles and industry whitepapers. The controls in the two lists are matched in Table-2.

Table 2: Comparison of ATT&CK mitigations and ZTA components

Mitigation suggested by MITRE ATT&CK (See Table-1)	Corresponding component of ZTA as suggested by NIST SP 800-27, academia and/or industry (See the list in the previous page)
Application Isolation and Sandboxing (1)	Application segmentation (5)
Exploit Protection (2)	Traffic filtering (8)
Network Segmentation (3)	Network segmentation (4)
Privileged Account Management (4)	Least privilege principle (15)
Update Software (5)	Operational and forensic analysis (10) Configuration management (12)
Vulnerability Scanning (6)	Operational and forensic analysis (10)
Disable or Remove Feature or Program (7)	Application execution control (9)
Limit Access to Resource Over Network (8)	Network access control (6) System access control (7)
Multi-factor Authentication (9)	Centralized authentication of users, devices, and applications (accompanied by SSO and MFA) (1)
User Account Control (10)	User account management (14) Least privilege principle (15)
Application Developer Guidance (11)	Guidance and training of application developers (16)
Password Policies (12)	Policy enforcement (11)
User Training (13)	Training and awareness (17)

All of the mitigations suggested by MITRE ATT&CK to prevent APT29 initial access have been matched by technologies, methods, and procedures recommended for zero trust implementations by NIST SP 800-207, academia, or industry.

NIST's SP 800-207 can be considered one of the principal resources in defining ZTA. It provides comprehensive guidance for organizations that plan to switch to ZTA. However, Vulnerability Scanning, Application Developer Guidance, and User Training have not been explicitly or strongly mentioned in NIST SP 800-207. Authors suggest that the future revision of NIST SP 800-207 should include these controls; even if they are not at the core of zero trust, they are essential preventative controls.

Three ZTA components in the list don't match any mitigations suggested by MITRE ATT&CK to prevent the APT29 threat actor. These components are (1) encryption, (2) continuous monitoring, and (3) data discovery, tracking, and analysis (Asset inventory). It is an anticipated situation; because the authors reviewed a number of academic and industry papers to extract ZTA components. It has been expected that the set of extracted components will be more extensive than the set of mitigations suggested by MITRE ATT&CK.

4. Conclusion

Information security controls can be classified as preventive and detective controls; some are preventive and detective, such as antivirus software. Prevention is ideal; however, detection is a must. As a matter of fact, one hundred percent prevention is impossible, making absolute security an impossible task. Eventually, this reality makes detection a mandatory process for organizations.

The two tactics of MITRE ATT&CK that come before "Initial Access" are "Reconnaissance" and "Resource Development". Cyber deterrence and cyber diplomacy might be effective in preventing Reconnaissance and Resource Development. Both tactics are difficult to mitigate by organizations, though, because the prevention methods for these tactics are implemented in a domain that is predominantly out of direct control of organizations. Moreover, these prevention methods might not be as effective as the prevention methods of Initial Access and subsequent tactics after the Initial Access. The detection of reconnaissance and resource development can be accomplished by cyber threat intelligence services. It is an effective method frequently used by organizations. Initial Access is the first tactic after Reconnaissance and Resource Development that an organization has direct control on the mitigation steps to prevent the techniques under this tactic. Effective prevention of Initial Access techniques is key to keeping the networks secure. In the SolarWinds breach, after attackers succeeded in the initial attack, they progressed successfully in the subsequent tactics such as Execution, Persistence, and Privilege Escalation.

Zero Trust Architecture is a promising approach to prevent Advanced Persistent Threats. There is a significant overlap between MITRE ATT&CK controls to prevent Initial Access and ZTA components and principles extracted from a diverse set of academic and industry papers. Merely focusing on detecting APTs has proven insufficient, and prevention should be considered at a higher priority than it has been in the past. ZTA is a structured way of accomplishing that goal.

ZTA can prevent the APT actors at the Initial Access and, therefore, improve the security significantly. In this regard, the 2020 United States Federal Government Data Breach caused by SolarWinds Orion and other vendor vulnerabilities could have been prevented if the attacked networks had used ZTA; in the worst case, the harm caused by the attacks might be far less compared to the actual case.

References

- Alevizos, L., Ta, V.T. and Hashem Eiza, M. (2021) 'Augmenting zero trust architecture to endpoints using blockchain: A STATE-OF-THE-ART review', *Security and Privacy* [Preprint]. doi:10.1002/spy2.191.
- Bajak, F. (2020) 'Hack may have exposed deep US secrets; damage yet unknown', *AP News*. Available at: <https://apnews.com/article/technology-hacking-coronavirus-pandemic-russia-350ae2fb2e513772a4dc4b7360b8175c> (Accessed: 18 November 2021).
- Banach, Z. (2021) 'Vulnerability scanning with PAM in zero trust environments'. Available at: <https://www.netsparker.com/blog/docs-and-faqs/vulnerability-scanning-privileged-access-management-zero-trust-environments/> (Accessed: 6 January 2022).
- Buck, C. et al. (2021) 'Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust', *Computers & Security*, 110, p. 102436. doi:10.1016/j.cose.2021.102436.

- Campbell, M. (2020) 'Beyond Zero Trust: Trust Is a Vulnerability', *Computer*, 53(10), pp. 110–113. doi:10.1109/MC.2020.3011081.
- Checkpoint (2021) 'The Ultimate Guide to Zero Trust Security'. Available at: <https://pages.checkpoint.com/the-ultimate-guide-to-zero-trust.html> (Accessed: 19 November 2021).
- Chen, L. *et al.* (2021) 'Research on the Security Protection Framework of Power Mobile Internet Services Based on Zero Trust', in *2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA). 2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, Kunming, China: IEEE, pp. 65–68. doi:10.1109/ICSGEA53208.2021.00021.
- Choi, S., Yun, J.-H. and Min, B.-G. (2021) 'Probabilistic Attack Sequence Generation and Execution Based on MITRE ATT&CK for ICS Datasets', in *Cyber Security Experimentation and Test Workshop. CSET '21: Cyber Security Experimentation and Test Workshop*, Virtual CA USA: ACM, pp. 41–48. doi:10.1145/3474718.3474722.
- Chowdhury, D.D. (2019) 'An Essential Guide to Zero Trust Security', p. 32.
- Cimpanu, C. (2020) 'SEC filings: SolarWinds says 18,000 customers were impacted by recent hack', *ZDNet*. Available at: <https://www.zdnet.com/article/sec-filings-solarwinds-says-18000-customers-are-impacted-by-recent-hack/> (Accessed: 18 November 2021).
- CISA (2020) 'CISA Issues Emergency Directive to Mitigate the Compromise of Solarwinds Orion Network Management Products'. Available at: <https://www.cisa.gov/news/2020/12/13/cisa-issues-emergency-directive-mitigate-compromise-solarwinds-orion-network> (Accessed: 18 November 2021).
- Cole, E. (2013) *Advanced Persistent Threat: Understanding the Danger and How to Protect Your Organization*. Syngress.
- Collier, Z.A. and Sarkis, J. (2021) 'The zero trust supply chain: Managing supply chain risk in the absence of trust', *International Journal of Production Research*, 59(11), pp. 3430–3445. doi:10.1080/00207543.2021.1884311.
- Colortokens (2021) 'The Definitive Guide to Zero-Trust Security'. Available at: <https://colortokens.com/ebook/definitive-guide-zero-trust-security/> (Accessed: 19 November 2021).
- Embrey, B. (2020) 'The top three factors driving zero trust adoption', *Computer Fraud & Security*, 2020(9), pp. 13–15. doi:10.1016/S1361-3723(20)30097-X.
- Garbis, J. and Chapman, J.W. (2021) *Zero Trust Security: An Enterprise Guide*. Berkeley, CA: Apress. doi:10.1007/978-1-4842-6702-8.
- Georgiadou, A., Mouzakitis, S. and Askounis, D. (2021) 'Assessing MITRE ATT&CK Risk Using a Cyber-Security Culture Framework', *Sensors*, 21(9), p. 3267. doi:10.3390/s21093267.
- Harding, L. and Sabbagh, D. (2020) 'Suspected Russian hackers spied on US federal agencies', *The Guardian*. Available at: <https://www.theguardian.com/world/2020/dec/14/suspected-russian-hackers-spied-on-us-federal-agencies> (Accessed: 18 November 2021).
- Horne, D. and Nair, S. (2021) 'Introducing Zero Trust by Design: Principles and Practice Beyond the Zero Trust Hype', p. 11.
- Hudson, B. (2014) 'Advanced Persistent Threats: Detection, Protection and Prevention'. Sophos.
- Karabacak, B. and Tatar, U. (2014) 'Strategies to Counter Cyberattacks: Cyber threats and Critical Infrastructure Protection', in Edwards, M. (ed.) *NATO Science for Peace and Security Series - E: Human and Societal Dynamics*, pp. 63–73.
- Manocha, H. *et al.* (2021) 'Security Assessment Rating Framework for Enterprises using MITRE ATT&CK® Matrix', p. 11.
- Mehraj, S. and Banday, M.T. (2020) 'Establishing a Zero Trust Strategy in Cloud Computing Environment', in *2020 International Conference on Computer Communication and Informatics (ICCCI). 2020 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, pp. 1–6. doi:10.1109/ICCCI48352.2020.9104214.
- Messaoud, B.I.D. *et al.* (2016) 'Advanced Persistent Threat: New analysis driven by life cycle phases and their challenges', in *2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS). 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS)*, Marrakesh, Morocco: IEEE, pp. 1–6. doi:10.1109/ACOSIS.2016.7843932.
- MITRE (2021a) 'APT29 Attack Navigator'. Available at: <https://mitre-attack.github.io/attack-navigator/#layerURL=https%3A%2F%2Fattack.mitre.org%2Fgroups%2FG0016%2FG0016-enterprise-layer.json>.
- MITRE (2021b) 'Frequently Asked Questions'. Available at: <https://attack.mitre.org/resources/faq/> (Accessed: 18 November 2021).
- Modderkolk, M. (2018) *Zero Trust Maturity Matters*. Master's Thesis. Utrecht University.
- Mohamed, N.A., Jantan, A. and Abiodun, O.I. (2018) 'An Improved Behaviour Specification to Stop Advanced Persistent Threat on Governments and Organizations Network', *Hong Kong*, p. 6.
- Moon, D. *et al.* (2014) 'MLDS: Multi-Layer Defense System for Preventing Advanced Persistent Threats', *Symmetry*, 6(4), pp. 997–1010. doi:10.3390/sym6040997.
- Nayak, S. (2021) 'What Are Advanced Persistent Threats (APTs) and How Can You Contain Them?' Available at: <https://colortokens.com/blog/advanced-persistent-threats-apt/> (Accessed: 19 November 2021).
- NSA (2021) 'Embracing a Zero Trust Security Model'.
- NSA, CISA and FBI (2021) 'Russian SVR Targets U.S. and Allied Networks'.
- Outkin, A.V. *et al.* (2021) 'Defender Policy Evaluation and Resource Allocation Using MITRE ATT&CK Evaluations Data', *arXiv:2107.04075 [cs]* [Preprint]. Available at: <http://arxiv.org/abs/2107.04075> (Accessed: 19 November 2021).
- Pell, R. *et al.* (2021) 'Towards Dynamic Threat Modelling in 5G Core Networks Based on MITRE ATT&CK', *arXiv:2108.11206 [cs]* [Preprint]. Available at: <http://arxiv.org/abs/2108.11206> (Accessed: 19 November 2021).

- Rivas, G. (2019) 'Zero Trust Model: An Effective Approach to Protecting Your Digital Environment', 14 January. Available at: <https://www.gb-advisors.com/the-zero-trust-model/> (Accessed: 6 January 2022).
- Rose, S. *et al.* (2020) *Zero Trust Architecture*. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-207.
- Sanders, G. *et al.* (2021) 'Integrating Zero Trust and DevSecOps'.
- da Silva, G.R., Macedo, D.F. and dos Santos, A.L. (2021) 'Zero Trust Access Control with Context-Aware and Behavior-Based Continuous Authentication for Smart Homes', in *Anais do XXI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg 2021)*. *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, Brasil: Sociedade Brasileira de Computação - SBC, pp. 43–56. doi:10.5753/sbseg.2021.17305.
- Splunk (2021) 'The Essential Guide to Zero Trust'. Available at: https://www.splunk.com/en_us/form/the-essential-guide-to-zero-trust.html (Accessed: 19 November 2021).
- Strom, B. *et al.* (2017) 'Finding Cyber Threats with ATT&CK-Based Analytics', (16), p. 53.
- Tao, Y., Lei, Z. and Ruxiang, P. (2018) 'Fine-grained Big Data Security Method Based on Zero Trust Model', in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. Singapore: IEEE, pp. 1040–1045.
- Tatar, U., Karabacak, B. and Gheorghe, A. (2016) 'An Assessment Model to Improve National Cyber Security Governance', in Zlateva, D.T. and Greiman, V.A. (eds) *11th International Conference on Cyber Warfare and Security*. ACPI Limited, pp. 312–319.
- Uttecht, K.D. (2020) *Zero Trust (ZT) Concepts for Federal Government Architectures*. Technical Report 1253. Lexington: Lincoln Laboratory.
- Van Driel, R. (2015) 'Zero Trust and APT'. Available at: <https://www.linkedin.com/pulse/zero-trust-apt-ruud-van-driel-cissp/>. (Accessed: 19 November 2021).
- Vanickis, R. *et al.* (2018) 'Access Control Policy Enforcement for Zero-Trust-Networking', in *2018 29th Irish Signals and Systems Conference (ISSC)*. *2018 29th Irish Signals and Systems Conference (ISSC)*, Belfast: IEEE, pp. 1–6. doi:10.1109/ISSC.2018.8585365.
- VMware (2022) 'Implement Zero Trust Security'. Available at: <https://www.vmware.com/solutions/zero-trust-security.html> (Accessed: 6 January 2022).
- Wikipedia (2021a) '2020 United States federal government data breach'. Available at: https://en.wikipedia.org/wiki/2020_United_States_federal_government_data_breach (Accessed: 18 November 2021).
- Wikipedia (2021b) 'Zero trust security model'. Available at: https://en.wikipedia.org/wiki/Zero_trust_security_model.
- Xiong, W. *et al.* (2021) 'Cyber security threat modeling based on the MITRE Enterprise ATT&CK Matrix', *Software and Systems Modeling* [Preprint]. doi:10.1007/s10270-021-00898-7.
- Zaheer, Z. *et al.* (2019) 'eZTrust: Network-Independent Zero-Trust Perimeterization for Microservices', in *Proceedings of the 2019 ACM Symposium on SDN Research. SOSR '19: Symposium on SDN Research*, San Jose CA USA: ACM, pp. 49–61. doi:10.1145/3314148.3314349.

Advancing cybersecurity capabilities for South African organisations through R&D

Zubeida Casmod Khan and Nenekazi Nokuthala Penelope Mkuzangwe

Council for Scientific and Industrial Research, Pretoria, South Africa

zdawood@csir.co.za

nmkuzangwe@csir.co.za

Abstract: There is a growth of cyber-attacks in South Africa. Seeing that there are over 38 million Internet users in South Africa, this is no surprise. The South African government has published the National Cybersecurity Policy Framework (NCPF) and Protection of Personal Information Act (POPIA) to move towards mitigating cyber threats due to the increase of the presence of South African organisations and citizens in cyber space. This demonstrates that there is a need for organisations to have a clear roadmap to implement and improve on their own cybersecurity capabilities. South African organisations need to take a proactive stance in cybersecurity because businesses rely heavily on technology for day-to-day operations. Currently cyber-attacks cost South African organisations over R2 billion, and the current work-from-home arrangement that most organisations have implemented will only worsen the situation. While a cybersecurity roadmap will differ in every organisation based on the organisation's vision, goals, and objectives, along with their information technology (IT) and operations technology (OT), a starting point is perhaps the identification of key research and development (R&D) areas together with key activities that organisations can focus on in order to improve their cybersecurity capabilities. Cybersecurity capabilities are tools that organisations use to strengthen their organisation and protect themselves from potential cyber threats. The purpose of this study was to investigate R&D areas that organisations should invest in for the purpose of improving their cybersecurity capabilities. There are various subfields in cybersecurity that can be explored for organisations to advance their cybersecurity capabilities. Five integral R&D dimensions were identified together with key activities and are presented and discussed. A conceptual framework is also presented which maps the R&D dimensions and activities to the main pillars of cybersecurity, i.e., People, Processes, and Technology. South African organisations could reference the framework and adapt it for their business needs to protect themselves against potential cyber threats.

Keywords: cybersecurity, cyber threat, cybersecurity capability, cyber-attack, research and development

1. Introduction

Since the COVID-19 pandemic, many professionals are working from home, which has resulted in the growth of cyber-attacks at an alarming rate. In South Africa, critical infrastructure such as its rail, port and pipeline company, Transnet, has been targeted in an act of cyber warfare (Slabbert and Peyper, 2021). Most organisations have systems with data that are connected to the Internet which opens avenues for cybercrime. The cybersecurity statistics of South Africa demonstrate that most organisations have not been able to protect their computer systems. Kaspersky announced that South Africa faced millions of cyber-attacks in 2020 (Kaspersky, 2020).

It is essential that organisations protect themselves from such threats as there is an increased presence in cyber space with cybersecurity. To do this, cybersecurity capabilities need to be developed or improved. The problem is that organisations are unclear on how to enhance their cybersecurity capabilities. The importance of enhancing cyber capabilities is emphasised by the Employment and Labour minister of South Africa, Thulas Nxes, stating that the South African government plans to also ramp up its cybersecurity capabilities following the cyber-attack on Transnet (Business Tech, 2021). While the South African government has promulgated the National Cybersecurity Policy Framework (NCPF) and Protection of Personal Information Act (POPIA), there has not been a national cybersecurity strategy, highlighting key research and development (R&D) activities that need to be investigated to strengthen cybersecurity capabilities.

To solve this problem of insufficient cybersecurity capabilities for organisations, the researchers aim to investigate the cybersecurity capabilities that other countries have invested in, and to categorise these capabilities in a theoretical framework, together with key activities. This is elucidated in a cybersecurity capability R&D framework.

The remainder of the paper is structured as follows. Section 3 provides an outline of work on cybersecurity R&D areas other countries focus on to strength their cybersecurity capabilities while Section 4 introduces the R&D dimensions which have been categorized. Section 5 presents the R&D framework, and the work is concluded in Section 6.

2. Related Works

This section presents existing literature on the cybersecurity research and development areas that are proposed to enhance cybersecurity capabilities in other countries. According to Di Franco (Franco, 2018), there is a need for multi-disciplinary research in modelling and designing future system in order to make them more easily understandable. This will assist with understanding how people adopt and use new technologies and how risks are perceived in digital space (Franco, 2018). The output of this research will address the rapid societal change and risks brought by the digital transformation (Franco, 2018). Another challenge recognised by the work (Franco, 2018) is those associated with the shortage of cybersecurity experts; this roots from issues at a curriculum level. Other noteworthy cybersecurity R&D areas recognised by the work of Di Franco (Franco, 2018) include Artificial Intelligence (AI), and Quantum Technology. AI techniques are expected to enhance cybersecurity by either automating certain routine tasks or assisting human system managers to monitor, analyse, detect and respond to adversarial threats to cyber systems (Institute for Information Infrastructure Protection, 2003) (Whitehouse, 2016). The impact of quantum technologies in security can be divided into those technologies that can defeat the existing data encryption systems and those that can provide secure communication that is resistant to quantum computing attacks (Franco, 2018). According to (ID Quantique, 2020), a technology that can protect from both classical and quantum computing attacks is well suited for protecting national security communication.

In addition to the aforementioned key R&D areas identified by Di Franco (Franco, 2018), the Federal Cybersecurity Research and Development Strategic Plan (Whitehouse, 2016) identifies trustworthy distributed digital infrastructure (TDDI), privacy, and secure hardware and software as priority points. TDDI is an important element in promoting the growth of new industries that will help drive a nation's economic growth (Whitehouse, 2016). TDDI will enable the distribution of computing to new network edge, acceleration of the roll-out of next generation telecommunication and information communications infrastructure such as Fifth-Generation and post Fifth-Generation wireless networks (Whitehouse, 2016). Cybersecurity aspects of privacy involves protecting the collection, disclosure, and use of an individual's private information, including identity; patterns of behaviour; and economic, social, or other discriminators (Whitehouse, 2016). Secure hardware and software enable the development of resilient and survivable cyber capabilities (Whitehouse, 2016).

Wireless networks include not only wireless telecommunications but a diverse set of end devices such as sensors, process controllers, and information appliances for home and business users and domain specific networks, such as vehicular communication networks and sensor networks; has been listed as one of the key research areas as there are many security concerns with it (Institute for Information Infrastructure Protection, 2003). Other R&D areas identified (Institute for Information Infrastructure Protection, 2003) include metrics and models, and Law, Policy, and Economic Issues, and digital forensics. Digital forensic assists with uncovering what happened during the cyber-attack, that is, how the incident happened, who did it, how to respond to it and how to stop it from happening in the future (DRS, 2018). This can help in building threat intelligence database, etc. (DRS, 2018). Therefore, (National Research Foundation: Singapore, 2021) suggests that applied research should be conducted for acquiring digital evidence and forensics in order to reduce the analysis time required to comprehensively locate and evaluate digital evidence in diverse, disperse and multi-tenanted systems.

Other cybersecurity measures that exist in literature include cyber-threat information and intelligence, network security protocols and cybersecurity governance. Cyber-threat information and intelligence is about indicators (system artifacts or observables associated with an attack), tactics, techniques, and procedures, security alerts, threat intelligence reports and recommended security tool configurations can help an organisation to identify, assess, monitor, and respond to cyber-threats (Johnson, Feldman and Witte, 2017). Where threat intelligence is threat information that has been aggregated, transformed, analysed, interpreted, or enriched to provide the necessary context for decision-making processes. Sharing cyber threat information with other organisations gives an organisation access to knowledge, experience and capabilities of the other organisations and helps the organisation gain a more complete understanding of the threats the organisation may face (Johnson, Feldman and Witte, 2017). Using this knowledge, an organisation can make threat-informed decisions regarding defensive capabilities, threat detection techniques, and mitigation strategies (Johnson, Feldman and Witte, 2017).

Network security protocols preserve the security and integrity of the data over a network. Different methodologies, techniques, and processes are involved in these protocols to secure the network data from any illegitimate attempt to review or extract actual data content (Reddy, 2020).

To establish and maintain effective cybersecurity governance program, (Swinton and Hedges, 2019) suggests that organizations should clearly define risk management policies, strategy, and goals and ensure that standardised processes are repeatable. Furthermore, (Swinton and Hedges, 2019) suggests that cybersecurity governance should be measurable and enforced so that there is accountability for compliance across all personnel levels. The organisations should ensure that senior leadership remains engaged for the lifecycle of the cybersecurity governance program and adequate resources are available to meet the organisation's basic cybersecurity governance and compliance needs (Swinton and Hedges, 2019).

3. Exploring R&D dimensions

Following the literature overview in the previous section, key R&D areas were identified and defined as dimensions. The selected R&D dimensions can be used to enhance organisations' cyber defence capabilities to address emerging cybersecurity risks. Recommendation activities are grouped together for each R&D dimension. The following sections provide the identified R&D areas to focus on.

3.1 Cybersecurity governance

This refers to the set of people, processes, and tools that are used to ensure that cyber risks are mitigated in an organisation. It is vital that organisations have a framework for how to deal with both attempted and successful cyber-attack. This framework must explain how an organisation can identify attacks, protect systems, detect and respond to threats, and recover from successful attacks. The framework can be used as a tool for aligning policies, business, and technological approaches to manage the cybersecurity risks (National Institute of Standards, 2014).

The framework also enables organisations to describe their current cybersecurity posture and target state of cybersecurity, identify, and prioritise opportunities for improvement within the context of a continuous and repeatable process, assess progress toward the target state and communicate cybersecurity risks to internal and external stakeholders (National Institute of Standards, 2014).

As organisations are moving towards the use automation technologies to enhance their functionality, this may introduce new/unknown operational risks and design security vulnerabilities associated with these technologies, which may lead to cybersecurity risks. This calls for research into the understanding of both operational, vulnerabilities, and cybersecurity risks associated with using these technologies. It further requires redefining cybersecurity governance in order to accommodate the handling of the risks associated with using automation technologies (Deloitte, 2019) . In order to establish and maintain effective cybersecurity governance the following R&D recommendations are made:

1. Research on how to develop a cybersecurity framework for managing cybersecurity risks (e.g., investigate best practice cybersecurity frameworks such as the NIST framework).
2. Research on risks associated with adopting automation technology, operational, design, and cybersecurity related frameworks.
3. Develop a cybersecurity framework that is aligned to the organisation's business drivers and security considerations that are specific to the use of the organisation's technology (both manual and automated).

3.2 Cybersecurity enabling techniques and technologies

Technologies, specifically those that enable good cybersecurity, need to be adapted for organisational use. In this R&D area we group together the following techniques and technologies: Artificial Intelligence (AI), Quantum Information Science (QIS), Trustworthy Distributed Digital Infrastructure (TDDI), secure hardware and software and vulnerability investigation technologies, secure communication, network protocols, wireless network security and remote, access security.

AI: AI can enhance cyber-attack deterrence, protection, detection, and response (Whitehouse, 2016). Furthermore, AI/ML can be used not only as a defensive measure in the context of cyber warfare but as an offensive tool also (Cummings, 2017). However, the adoption of AI is subjected to the challenges which include the reasoning and outcomes of AI algorithms/techniques being not understandable which may lead to incorrect

decisions, attacks on AI algorithms, etc (Institute for Information Infrastructure Protection, 2003). Therefore, the following R&D recommendations are made:

- Investigate the AI based technologies that can be used in the context of cybersecurity/cyber warfare (cyber-attack monitoring, deterrence, protection, detection, response, and as an offensive tool).
- Invest on meeting the requirements associated with the deployment of these technologies that include obtaining big datasets for training and testing the technologies, high computing machines for the implementation of these technologies and big data storage that is secure.
- Research on gaining understanding of attacks against these technologies and tools to protect these technologies from those attacks (Whitehouse, 2016).
- Research onto the risks associated with adopting these technologies and how they can be mitigated.
- Investigate legal and ethical issues associated with employing AI in cyber warfare.
- Implement AI technologies to monitor, analyse, detect, and respond to adversarial threats to cyber systems. This is typically used for external threats.
- Develop methodologies to validate and interpret results from AI systems against human perceptions and expectations (Institute for Information Infrastructure Protection, 2003).
- Develop definitions, models, and metrics of security and trust that can be used to evaluate AI cybersecurity systems and AI-based cybersecurity controls (Whitehouse, 2016).

QIS: QIS can enhance protection from cyber-attacks (Whitehouse, 2016). Quantum technologies can provide secure communication that is resistant to quantum computing attacks and promise to achieve a higher level of security than classical computing limits (Whitehouse, 2016). However, it can also be used to defeat existing data encryption systems (Whitehouse, 2016). Therefore, understanding quantum capabilities and finding ways to overcome their fundamental limits is vital (Whitehouse, 2016). The following R&D recommendations are made:

- Conducting research to gain understanding of which quantum technologies exist, how they can be used to enhance cybersecurity and their vulnerabilities and adoption.
- Investigate the deployment requirements of such technologies.
- Investigate the cybersecurity risks associated with using such technologies and how they can be mitigated.
- If such home-grown technologies cannot be obtained, partner with home companies/organisations that can conduct R&D of those technologies, their deployment requirements and cybersecurity risks and mitigations.

TDDI: TDDI assists in facilitating next-generation telecommunications and information communications infrastructure such as 5G wireless network, IoT etc (Institute for Information Infrastructure Protection, 2003). These technologies are part of the fourth industrial revolution (4IR) technologies and since the 4IR technologies are being increasingly used, it is at the best interest of organisations and countries to start preparing for the adoption of these technologies. Therefore, in order to prepare for the adoption of these technologies, the following R&D recommendations are made:

1. The starting point for adoption of the relevant 4IR technologies is to conduct research to gain understanding of the distributed digital infrastructure that is needed to enable the deployment of the relevant 4IR technologies.
2. Research into gaining understanding of which 4IR technologies are relevant to the organisation's business.
3. Research to gain understanding of the operation and deployment of the infrastructure and technologies.
4. Research to gain understanding of security risks that are associated with the infrastructure and technologies and ways to mitigate those risks.

Secure Hardware and Software and Vulnerability Investigation Technologies: Secure hardware and software to enhance the protective and detective aspect of cybersecurity (Whitehouse, 2016). However, hardware, firmware, and software tend to have design defects that present security vulnerabilities (Whitehouse, 2016). Therefore, the following research recommendations are made:

1. Develop mechanisms and tools that verify the security properties of hardware before its deployed (Institute for Information Infrastructure Protection, 2003).
2. The development of security solutions to patch vulnerabilities and prevent unauthorised program modification in order to protect against future attacks.
3. Investigate technologies for investigating system vulnerabilities, discovering unauthorised hardware modifications and unauthorised programs and obtain them through R&D if not available (ATLA, 2020).

4. Obtain/develop system vulnerability investigation personnel.
5. Conduct research to gain understanding of penetration testing methods and their implementation or recruit penetration testing professionals to implement penetration testing or train some employees to become penetration testers through courses/workshops.

Network Security Protocols: Network security protocols play an important part in protecting network traffic both in wired and wireless networks. These protocols use encryption and authentication to protect these networks. Traditional encryption algorithms tend to be heavy for wireless devices such as sensors (ID Quantique, 2020). Encryption can be broken by quantum computing attacks. This means, there is a need for developing encryption/data protection algorithms/techniques that are suitable for use on wireless devices and resistant to quantum computing attacks. Therefore, the following recommendations are made in order to enhance network security protocols functionality:

- Investigate and develop quantum computing resistant encryption algorithms that can be used in the protocols.
- Investigate and develop encryption algorithms or data protection algorithms that are appropriate to be used with wireless network devices.

Wireless Network Security: The adoption of wireless network technologies/devices is growing among organisations. According to (Institute for Information Infrastructure Protection, 2003), some of these devices tend to have limited power, processing, storage etc., that limit their security capabilities. Furthermore, security solutions for wired networks may not be usable in wireless networks due to the unfair differences between these network systems capabilities (Institute for Information Infrastructure Protection, 2003). Therefore, the following research recommendations are made in order to enhance wireless security capability of an organization:

- Research and develop security solutions for an organisation's wireless networks that include vulnerability investigation of their wireless systems.
- Investigate traffic monitoring and analysis technologies to gain an understanding of the state of the network.

3.3 Cybersecurity capacity building

To put cybersecurity into practice, an organisation needs experts on the topic, whether they are hired or outsourced. To achieve this capacity in terms of cybersecurity experts, the following R&D recommendations are made:

- Hire or outsource diverse postgraduate graduates and computer science researchers.
- Hire or outsource employees with cybersecurity skills and certifications based on the gaps determined during the skills audit.
- Identify private and public entities that can enable the development of novel research and technology.
- Identify cybersecurity institutions for training staff members.

3.4 Cyber threat intelligence

This is of the utmost importance to determine the components of the organisation's software system that needs to be protected and the type of security threats to protect the component against attacks. Cyber-attacks are continuously increasing in frequency and sophistication and this presents significant challenges for organisations that must defend their data and systems from capable threat actors (Johnson *et al.*, 2016).

Joining cyber threat information and intelligence sharing communities helps the organisation to gain understanding of threats that it can face. The knowledge obtained from sharing can be used to make threat-informed decisions regarding defence capabilities, threat detection techniques and mitigation strategies (Johnson *et al.*, 2016). Researchers recommends the use of standardised data representation formats (e.g., Structured Threat Information Expression (STIX)) and transport protocols (e.g., Trusted Automated Exchange of Intelligence Information (TAXII)) to make the sharing of cyber-threat information and intelligence easier (Johnson *et al.*, 2016). Data representation formats provide a reporting structure and language that can be used to capture descriptions of threats. Transport protocols are mechanisms/protocols for the transportation of threat information/intelligence.

Organisations need to join cyber-threat information and intelligence sharing communities and invest in technologies that enable cyber-threat information and intelligence sharing and the personnel to operate these technologies. Recommendations for keeping up to date with existing and new cyber threats are as follows:

- Determining the components of a system that needs to be protected and the type of security threats to protect the component against.
- Implementing cyber threat analysis training.
- Investigate existing local and international cyber threat information sharing communities.
- Joining cyber-threat information and intelligence sharing communities (locally, South African Cyber Hub and internationally) that are of relevance to the domain of the business.
- Obtaining technologies that enable automated cyber-threat information and intelligence sharing.
- Investigating and adopting the data representation formats and transport protocols used by these sharing communities.
- Securing personnel to work/operate these technologies. The adoption of standardised data formats and transport protocol will require a capable workforce.

3.5 Cybersecurity awareness building

People are at the centre of successful cybersecurity at an organisation. However, people typically think of cybersecurity as an isolated IT office problem, or something that is difficult to achieve (Institute for Information Infrastructure Protection, 2003). This indicates that there is a need to empower employees/citizens with the knowledge and a sense of shared responsibility to practice safe and informed behaviours in the cyberspace (Franco, 2018). At an organisation level, the following recommendations are made in order to educate employees to practice safe and informed behaviours in the cyberspace:

- Conduct cybercrime awareness and prevention campaigns. Enforce these campaigns and training at all levels of an organisation from top to bottom level, i.e., from the CEO to the intern level.
- Conduct periodic training across the organisation using a combination of eLearning, workshops, policy rollouts, simulated phishing emails, blogs, social media posts, articles, risk assessments, and surveys.
- Participation in cyber games by employees. According to (Awojana and Chou, 2019), game based learning allows learners to actively learn and practise the right way things should be done, while having fun. This approach leads to effectiveness on cybersecurity awareness training and practical skill acquisition for learners from diverse backgrounds.
- Investigating local and international cyber conferences for attendance and participation. Invest in allowing employees to attend cyber conferences such as Black Hat¹.
- Identify organisations to develop technologies and tools for cybersecurity awareness.

4. R&D Framework

The dimensions with tasks explored in the aforementioned section are used to drive the R&D framework. The framework, shown in Figure 1, is built upon the three pillars of cybersecurity, i.e., People, Processes, and Technology (PPT) which is harmonised to deliver a holistic approach to drive cybersecurity R&D.

People is all about the personnel involved in cybersecurity R&D. This involves employees, training institutions, communication mechanisms etc. Processes encompasses the rules, regulations, frameworks, campaigns etc. that the R&D dimensions relate to, while technology involves the systems, applications, hardware, software, etc. that are relate to the R&D. Each pillar of PPT are closely related to each other and each have an important role to play for organisations' cyber security. Organisations are encouraged to look at the pillars, i.e., PPT and implement the key activities for each pillar to ensure adequate security.

In the following sections, each pillar is introduced, together with its goal, and key activities. The key activities are devised from summarising the R&D recommendations from Section 4 and mapping it to a pillar (i.e., PPT).

4.1 People

Every employee of an organisation has a role to play in terms of cyber security. People need to partake in several tasks and activities to ensure effective cyber security of an organisation. Thus, the main goal for this pillar is: *to prepare people in an organisation by training and other human-centric, interactive activities to ensure that an organisation's IT systems are secure*. Therefore, this pillar is linked to the **cybersecurity awareness, cyber threat**

¹ <https://www.blackhat.com/us-21/>

intelligence, and **cybersecurity capacity building** R&D dimensions of the framework. The activities associated with these R&D dimensions that can be adopted in this pillar are as follows:

1. Identify organisations to provide tools and technologies for cyber security awareness training.
2. Invest in and implement cyber security awareness training, cyber security campaigns and game participation.
3. Identify private and public institutions to collaborate with the organisation to build capacity.
4. Identify cyber security training institutions where the personnel can be trained or up skilled since training is an on-going activity as cyber security progresses and criminals use sophisticated and novel attacks to gain access to organisations.
5. Invest in and implement cyber threat intelligence training or secure personnel to perform cyber threat intelligence. This includes hiring computer science graduates as well as hiring experience individuals and outsourcing if needed.

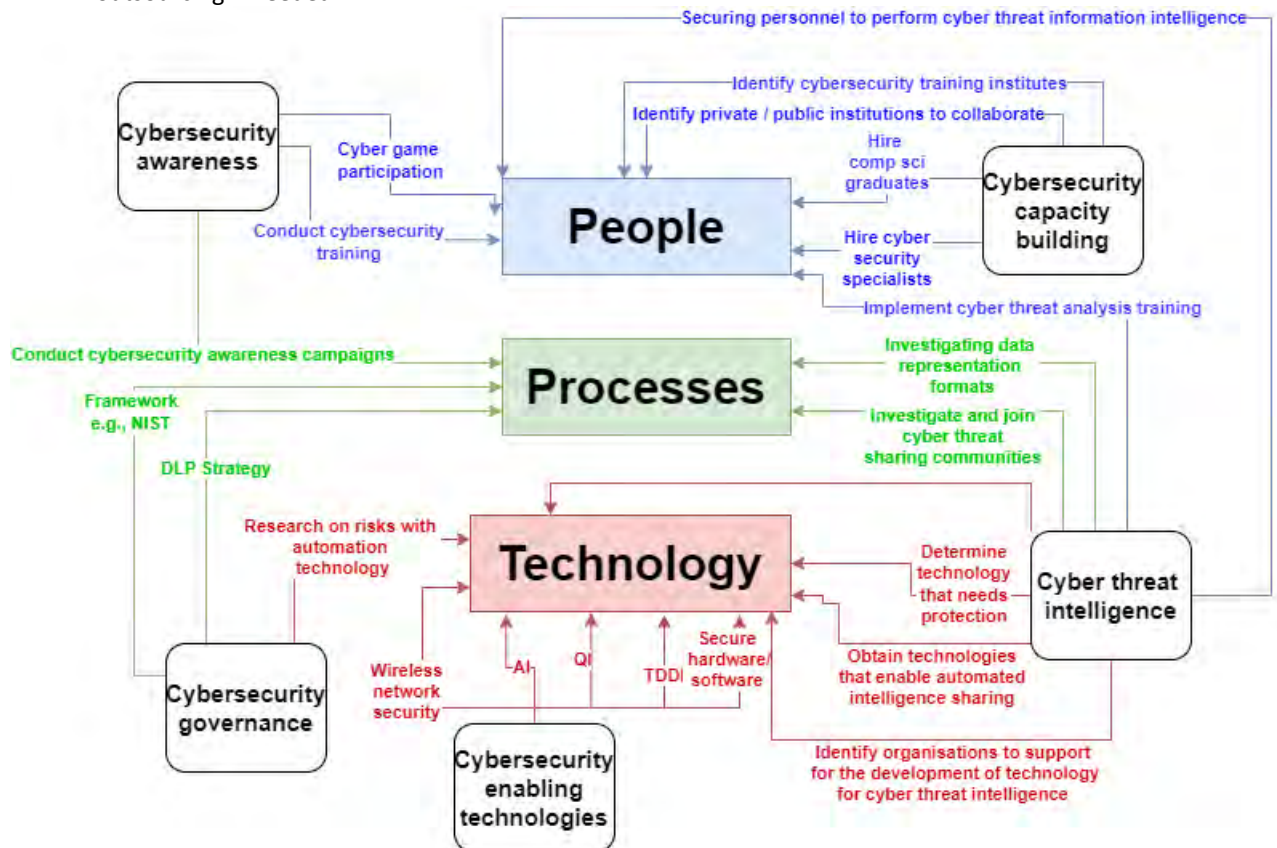


Figure 1: R&D framework for cybersecurity capabilities for organisations.

4.2 Processes

Processes are steps that need to be taken to ensure cyber security and protection of an organisation. This involves the policies, procedures and other practises that should be taken by an organisation. The main goal that organisations should seek for this pillar is: *"to continuously adapt and improve information security processes to ensure adequate response to security incidents and reduce cyber risk."* This pillar is linked to **cybersecurity governance**, and **cyber threat intelligence** R&D dimensions of the framework.

The activities associated with these R&D dimensions that can be adopted in this pillar are as follows:

1. The development and implementation of cyber security awareness training, cyber security campaigns and game participation plans.
2. The development and implementation of cyber security frameworks e.g., NIST.
3. Implement DLP processes to protect sensitive information of the organisation.
4. Investigate and join cyber threat sharing communities that can be joined to share intelligence.
5. Investigate data representation formats and transport protocols used by cyber threat intelligence sharing communities.

4.3 Technology

Technology is used by people to automate processes to enforce cyber security for an organisation. The main goal that organisations should seek for this pillar is: “to utilise cyber security technologies to ensure the confidentiality, integrity, and availability of organisational data.” This pillar is linked to **cybersecurity governance**, **cyber enabling technologies** and **cyber threat intelligence** R&D dimensions of the framework.

The activities associated with these R&D dimensions that can be adopted in this pillar are as follows:

1. To perform research on risks associated with adopting automation technology.
2. To implement AI technologies to monitor, analyse, detect and respond to adversarial threats to cyber systems.
3. To implement technologies that enables information sharing for cyber security threats.
4. To use secure hardware and secure software to enhance the security of the organisation’s technology.
5. To determine the technology that needs to be protected.
6. TDDI and QIS are not widely used in most information systems yet, but organisations should monitor technology for when there are operational as it will change the nature of cyberattacks.
7. Implement software security solutions for the organisations’ wireless networks.
8. Identify private and public institutions that can enable and support for the development of research and technology for cyber threat intelligence tailored to the organisation at hand.

5. Conclusion

The increase in cybersecurity breaches in South Africa demands for organisations to equip themselves for protection. While the South African government has recently promulgated the NCPF and POPIA, a national strategy has not been published. This could cause organisations to be indifferent to cybersecurity measures that need to be taken to protect themselves. To solve this problem, the researchers have explored cybersecurity R&D areas of other countries and categorised these as dimensions with task recommendations. This has been conceptualised and the researchers presented a R&D framework for cybersecurity capabilities built upon the PPT pillars of cybersecurity. Organisations can now adopt this R&D framework for business needs towards achieving cybersecurity.

The proposed framework opens a plethora of avenues for future research. It would be beneficial to create strategies for certain types of organisations, i.e., the telecommunications industry, electricity industry etc. which would enable various sectors to further protect themselves.

References

- ATLA (2020) *R&D Vision Toward the Realization of a Multi-Domain Defense Force and Beyond Explanatory Documentation*. Available at: https://www.mod.go.jp/atla/en/policy/pdf/rd_vision_full.pdf (Accessed: 17 February 2021).
- Awojana, T. and Chou, T.-S. (2019) ‘Overview of Learning Cybersecurity Through Game Based Systems’, in *Proceeding of the 2019 Conference for Industry and Education Collaboration, CIEC 2019*. New Orleans, LA: Advances in Engineering Education.
- Business Tech (2021) ‘South Africa building database of security specialists as it recovers from major cyber attack’, *Businesstech*, August. Available at: <https://businesstech.co.za/news/it-services/510640/south-africa-building-database-of-security-specialists-as-it-recovers-from-major-cyber-attack/> (Accessed: 23 November 2021).
- Cummings, M. ‘Missy’ L. (2017) *Artificial Intelligence and the Future of Warfare*. Available at: <https://www.chathamhouse.org/publication/artificial-intelligence-and-future-warfare> (Accessed: 11 October 2018).
- Deloitte (2019) *Future of Risk in the Digital Era | Deloitte US*. Available at: <https://www2.deloitte.com/us/en/pages/advisory/articles/risk-in-the-digital-era.html> (Accessed: 19 November 2021).
- DRS (2018) *The importance of digital forensics in cyber security*, DRS. Available at: <https://www.drs.co.za/the-importance-of-digital-forensics-in-cyber-security/> (Accessed: 22 February 2021).
- Franco, F. Di (2018) *Analysis of the European R&D priorities in cybersecurity Strategic priorities in cybersecurity for a safer Europe*. Available at: https://www.enisa.europa.eu/publications/analysis-of-the-european-r-d-priorities-in-cybersecurity/at_download/fullReport.
- ID Quantique (2020) *The impact of quantum technology on national security - ID Quantique, ID Quantique*. Available at: <https://www.idquantique.com/the-impact-of-quantum-technology-on-national-security/> (Accessed: 17 February 2021).
- Institute for Information Infrastructure Protection (2003) *CYBER SECURITY RESEARCH AND DEVELOPMENT AGENDA*. Available at: https://web.stanford.edu/class/msande91si/www-spr04/readings/week8/2003_Cyber_Security_RD_Agenda.pdf (Accessed: 22 February 2021).
- Johnson, C. et al. (2016) *NIST Special Publication 800-150 Guide to Cyber Threat Information Sharing*. doi: 10.6028/NIST.SP.800-150.

- Johnson, C., Feldman, L. and Witte, G. (2017) *ITL BULLETIN FOR MAY 2017 CYBER-THREAT INTELLIGENCE AND INFORMATION SHARING*. Available at: <https://csrc.nist.gov/CSRC/media/Publications/Shared/documents/itl-bulletin/itlbul2017-05.pdf> (Accessed: 22 February 2021).
- Kaspersky (2020) 'South Africa, Kenya and Nigeria saw Millions of Cyber Attacks in 2020 and the Year is not over yet', *Africa newsroom*, September. Available at: <https://kaspersky.africa-newsroom.com/press/south-africa-kenya-and-nigeria-saw-millions-of-cyber-attacks-in-2020-and-the-year-is-not-over-yet?lang=en> (Accessed: 23 November 2021).
- National Institute of Standards (2014) *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*. doi: 10.6028/NIST.CSWP.04162018.
- National Research Foundation: Singapore (2021) *National Cybersecurity R&D Programme*. Available at: <https://www.nrf.gov.sg/programmes/national-cybersecurity-r-d-programme> (Accessed: 17 February 2021).
- Reddy, P. (2020) *Network Protocols and Its Security*, *Medium*. Available at: <https://priya-reddy.medium.com/network-protocols-and-its-security-47d68f356666> (Accessed: 22 February 2021).
- Slabbert, A. and Peyper, L. (2021) 'Transnet attack is cyber warfare', *Citypress*, 1 August. Available at: <https://www.news24.com/citypress/business/transnet-attack-is-cyber-warfare-20210801> (Accessed: 23 November 2021).
- Swinton, S. and Hedges, S. (2019) *Cybersecurity Governance, Part 1: 5 Fundamental Challenges*. Available at: <https://insights.sei.cmu.edu/blog/cybersecurity-governance-part-1-5-fundamental-challenges/> (Accessed: 19 November 2021).
- Whitehouse (2016) *FEDERAL CYBERSECURITY RESEARCH AND DEVELOPMENT STRATEGIC PLAN*. Available at: <https://www.nitrd.gov/groups/csia/>. (Accessed: 17 February 2021).

Zero Trust Container Architecture (ZTCA): A Framework for Applying Zero Trust Principals to Docker Containers

Darragh Leahy and Christina Thorpe

Technological University Dublin, Blanchardstown, Dublin, Ireland

ZeroTCA@gmail.com

christina.thorpe@tudublin.ie

Abstract: Containerisation is quickly becoming an accepted industry standard for development environments and Gartner, in a recent market forecast, estimated that by 2022 more than 75% of organisations will be using containers in production deployments. With this explosion in growth comes an added focus on security and best practices for using containers. The use of containers, in particular Docker containers, has altered some of the more traditional deployment paradigms by giving control of deployments to the development teams. This has massively benefited the DevOps release cycle, but at the expense of many mature security and review processes that are integrated into traditional deployments. Like all systems, containers need frameworks to guide best practices for deployments and to ensure mistakes are not made that increase the risk level or attack surface of an application or service using containers, or the containers themselves. Indeed, according to a recent presentation during DevSecCon24 by Justin Cormack, Security Lead at Docker Inc., Cormack believes most security issues related to Docker are due to misconfiguration rather than direct exploit. While work has been previously conducted with regards to container security and separately applying Zero Trust Networking Architecture to containers, in this work we will investigate the security state of a default deployment of the Docker container engine on Linux and analyse how the principals of Zero Trust Architecture can be extended beyond the domain of networking, distilled into a "Zero Trust Containers Architecture" and applied to secure Docker deployments. In order to determine this, research was conducted into the current state of Docker security and Zero Trust Architecture. Practical and theoretical attacks were reviewed against a default Docker deployment to identify common themes and areas of issue. Results were used to advise a generalised trust-based framework which was then used to analyse a Docker deployment and validate mitigation of a selection of the identified attacks, proving out the concept of the proposed "Zero Trust Container Architecture" framework.

Keywords: Zero-Trust Architecture, Containerisation, Virtualisation, Security, DevSecOps

1. Introduction

Regardless of the forecast advances made by Docker Inc. in securing the product, the current security state of Docker is still a concern. Many security issues do exist currently, this is driven by a combination of technical factors such as vulnerabilities in the Docker Engine and Linux OS, misconfigurations of the Docker system or limitations of its design, however equally important are the concepts of 'system espoused' versus the 'system in use' problem. That is, the difference between what the product developer aims to achieve and consequently what they design for, as opposed to how the system is actually used in the real world. Work in (Martin 2018) demonstrated that a fundamental difference exists in what Docker considers the recommended use case versus what is actually being done in the real world. These differences, combined with the technical issues, as well as the rapid adoption of Docker have combined to create a potential storm of security concern. Zero Trust Networking Architecture (ZTA/ZTNA) is a design paradigm for networks which identified the need to move the defensive perimeter down from the edge of the network. In a traditional edge firewall and Intrusion Prevention/Intrusion Detection system, the edge of the network was considered the point of security, with anything inside broadly trusted. ZTA advised to redefine this boundary away from the perimeter and down to the user and resource in use. ZTA re-frames the question of network security into one of trust, that is, the inherent insecurity of assumed trust that exists once inside the network perimeter. ZTA advises to assume no explicit trust within a system, that internal users accessing internal services without explicit authentication and authorisation cannot be trusted any more than external actors. ZTA's ethos can be summed up with the line 'never trust, always verify'. ZTA offers opportunities to be re-purposed for use in securing Docker containers and thereby mitigating some of the known configuration issues as well as minimise the surface area of exploits and vulnerabilities.

This research will review and identify the issues with a default deployment of Docker through a combination of practical experimentation to answer the core research question: ***Using a novel generalised trust-based framework, based on the proven principals of ZTA, can a Docker deployment be secured?*** Some ancillary questions have also been defined: (1) What issues currently affect a default deployment of Docker? (2) What

additional common misconfigurations affect a Docker deployment? (3) Can ZTA, or an abstraction thereof be applied to help secure a Docker deployment?

Using the strategic thinking and principles underpinning ZTNA, this paper aims to distil a trust-based framework that considers multiple aspects of a Docker container deployment. With the aim of evolving ZTA beyond just the scope of networking. To the best of our knowledge, the re-purposing of ZTA into a generalised trust framework and applying the results to containers has not been done before. The success of this trust-based framework will be validated through practical example and process as well as theoretical example. This is to be achieved by applying the investigated attacks and issues against various Docker use cases reviewed and secured using the new trust-based framework and reasoning out the improved security surface. This paper will focus on the most straightforward deployment of Docker that can be configured, which is also likely to be one of the most common real-world deployment scenarios. To this, we will propose and apply a trust-based framework adapted from ZTA.

2. Background

2.1 Containers



Figure 1 (a): Docker Container versus VM

Containers, as defined by Docker Inc., provide “a standard unit of software that packages up code and all its dependencies, so the application runs quickly and reliably from one computing environment to another”. While not all containerisation systems share the same ethos, the fundamental structures, as seen in Figure 1, are broadly similar. Where Virtual Machines (VMs) run on top of a hypervisor ‘multiplexing’ the hardware, containers run on top of the Host Operating System (Host OS) sharing and ‘multiplexing’ the kernel. Containerisation offers an alternate solution to full VMs, with the ability to address specific demands for simplicity and speed arising from modern development concepts like microservices. Containers present the perfect solution by offering rapid and idempotent deployment of container images from private or shared repositories, extremely fast boot times and portability. Figure 2 shows an example of just how fast the start time of a Docker container is relative to alternate solutions.

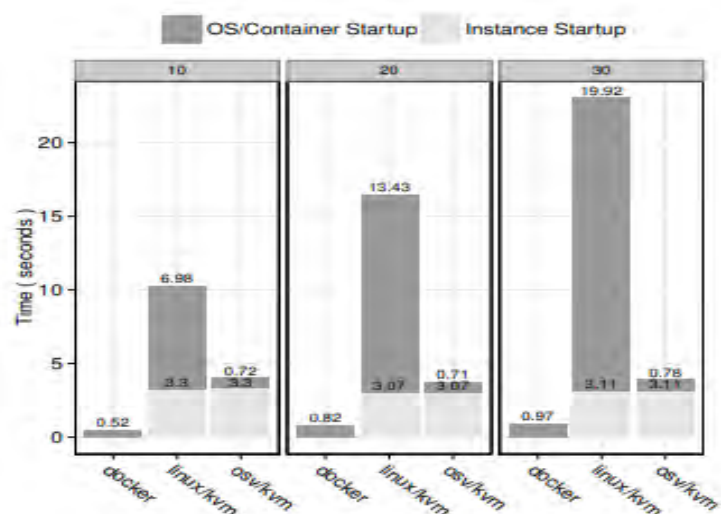


Figure 2: Instance and Operating System/Container Start-up for 10, 20 and 30 instances (Xavier et al., 2016)

2.2 Docker Containers

Docker as a container solution emerged in 2013. Docker views containers as a mechanism for packing specific applications inside a (relatively) idempotent image that can be shared and deployed but eschews the typical features of an OS like system services, daemons, etc. That makes Docker and similar projects much closer to a software distribution mechanism than a machine management tool.” (Graber 2016). Another important element of Docker that sets it apart from many other containerisation engines is the sheer ubiquity of adoption, with Docker claiming that container images have been downloaded from the Docker Hub over 130 billion times and that the Docker engine itself is responsible for sparking the ‘Containerization Movement’. Docker is one of the most recognisable names in containerisation. This, combined with its developer targeted approach in supporting rapid application development and deployment, has made it the focus of this paper.

2.3 Lifecycle of a Docker Container

Docker container images are built from a reference spec called a ‘Dockerfile’ (See Figure 3). A Dockerfile is commonly a yaml file (yaml is used by lots of different applications for both configuration and for defining data in a human-readable structured data format) that describes the steps required to assemble an image, allowing developers to select a base OS image, install dependencies and inject a copy of their software. Each of the steps in this file is converted into a ‘layer’ by the Docker engine with each layer combining to form the final container image. This image then becomes idempotent, such that while individual running instances can be changed, the image itself cannot be, allowing for ‘known good’ configurations, leading to extreme portability. These images are then pushed to an image repository, either private or shared, e.g., Docker Hub (also supports private use cases). Other developers, Ops teams, or indeed any permitted end user can then pull these images to their Docker deployments and have a container running the image in a highly repeatable, straightforward fashion. Rather than upgrading the container OS or application, when a new container version is released, the new container image can simply be download and the old image discarded.

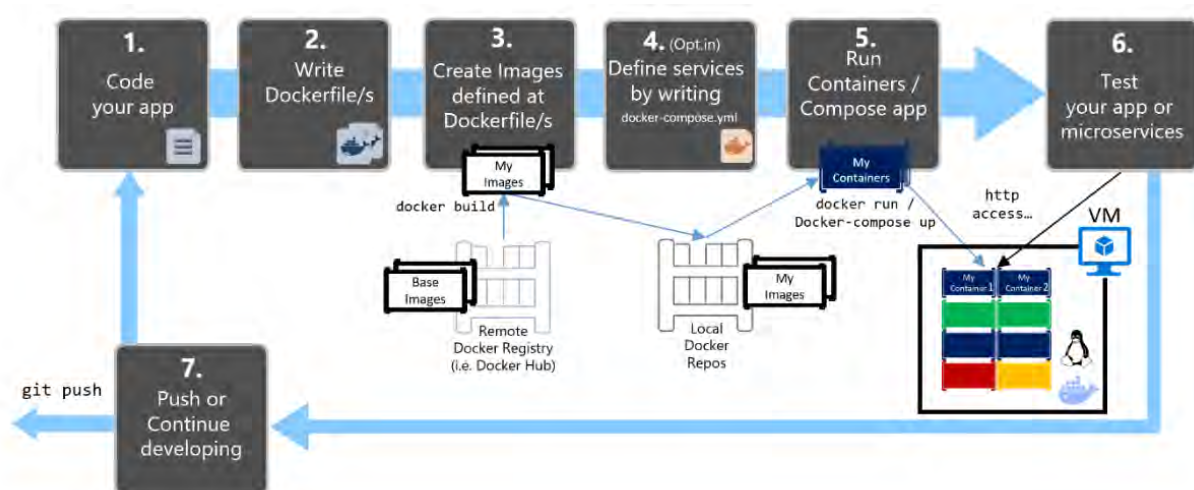


Figure 3: Step-by-step workflow for developing Docker containerized apps ¹

2.4 Zero Trust Network Architecture

Traditional network security generally focused on a perimeter-based model, implementing the majority of security systems at the network edge. This edge was typically secured with firewalls and included Intrusion Prevention/Intrusion Detection systems (Kerman 2020). Once inside the network edge however trust was broadly assumed, resulting in poor, if any, security systems policing and protecting the internal network environment. Zero Trust Network Architecture (ZTA, or ZTNA), exists as a set of security principals for securing networks. Originally proposed in (Kindervag, 2010), ZTA assumes that implicit trust itself is a vulnerability in the network environment and seeks to eliminate the risks posed by malicious actors, internal or external, abusing this implicit trust. This is achieved by removing the concept of a trusted internal network and untrusted external network, rather instead assuming all networks are untrusted and ensuring that devices and users within a network operated from a least permission model. This was to be further augmented by the visibility and auditing

¹ <https://docs.microsoft.com/en-us/dotnet/architecture/microservices/docker-application-development-process/docker-app-development-workflow>

of all activities taking place within a network. Kindervag proposed three key concepts for ZTA (Kindervag, 2010): (1) Ensure that all resources are accessed securely regardless of location (2) Adopt a least privilege strategy and strictly enforce access control (3) Inspect and log all traffic. ZTA is often boiled down to the following mantra: Never trust, always verify.

3. Related Work

Much work has been undertaken in the areas of container security, including Docker, and separately Zero Trust Architecture. Both topics being well established and studied, however there have been relatively few papers discussing Zero Trust Architecture for containers. Indeed, those papers that do exist in the crossover between container security and ZTA are focused primarily on implementing the traditional Zero Trust Networking approach to containers rather than abstracting the rules of ZTA into a more comprehensive framework to consider and govern the entire surface area of a container deployment, or more specifically as it relates to this paper, the surface area of a Docker container deployment.

(Surantha 2020) provides arguments for the use of ZTA from a technical standpoint in securing a container deployment, without evolving the concept of ZTA beyond the network. (de Weeve 2020) provides a more in-depth look at technical details, strategy and considerations when using ZTA with containers and identifies ZTA as more than purely a technical concept, but a security approach. (Yasrab 2018) provides a review of some of the most common issues affecting Docker containers and suggests several technical mitigations to address specific issues without presenting a framework to be used in addressing and mitigating issues going forward. (Martin 2018) conducted an extensive study of container security with practical demonstration from the standpoint of what Docker recommends vs what is seen in the wild and being done by major Cloud Service Providers. While this is an excellent paper that provides the basis of a mechanism for assessing Docker from a security perspective, it is driven from a very technical focus and does not provide a complete framework for analysing Docker deployments. (Bui 2015) contends that with proper security measures and lockdown, such as a reduction of the Linux CAPs assigned to a container and implementation of SELinux or AppArmor, as well as the removal of root use and privileged mode from containers, that containers could be considered relatively secure. Bui however does not present any framework or methodology that can be used to analyse a Docker deployment. (Rose 2019) provides a more comprehensive abstraction of ZTA, broadly arguing that ZTA can be used as a tool for strategic thinking, focusing much more on the concepts of trust and how this can be established rather than specific technical implementations.

4. Problem

In a 2020 Cyberthreat Defence Report published by CyberEdge, containers were listed as the number 1 weakest link in the security landscape. Containers are subject to bugs and vulnerabilities and therefore open to abuse. (Sultan 2019) states that containers are less secure than traditional VMs, and notes that this relative insecurity, or perceived insecurity, of containers remains one of the largest barriers to the ongoing adoption of containers. What makes Docker in particular an interesting target in the world of containers is threefold. First, we argue that the default state of a Docker deployment does not apply security best practices. Second, common real-world misconfigurations of Docker deployments increase the vulnerability and attack surface of a given deployment. Finally, Docker's large uptake by the developer community to assist in rapid development and deployment has resulted in the bypass of mature deployment processes. These factors have all combined to escalate the threat posed by Docker when deployed without the use of a methodology, or framework, for considering its security. In an article published on BlackHat, a leading security event series, Anthony Bettini, founder of Flawcheck, dedicates an entire section to vulnerabilities not necessarily needing a Common Vulnerabilities and Exposures (CVE) or code-based exploitation but rather those which attackers can take advantage of due to "implementation weaknesses (ultimately, design imperfections)" (Bettini 2015). This is further borne out by several other articles describing some of these real-world exploits or proving out the issue directly (Simonovich 2019), (Higashi 2018).

A similar investigation was conducted in 2019, that detected over 20,000 Docker instances, and a similar number of Kubernetes instances, via Shodan (Quist, 2019). Quist notes that while these 40,000 combined instances are not inherently vulnerable, they are at least demonstrating security misconfiguration in their deployments. The research did however discover a number of vulnerable databases and deployments leaking personal information. TrendMicro, ran a study to determine some real-world exploit scenarios against containers (Oliveira 2019). They discovered that malicious actors were using automated scanning to find exposed Docker APIs and

deploy cryptomining container images against them, as well as containers to run scans for other exposed hosts. An exposed Docker API is not the only issue, if we assume that a container has been accessed or exploited (via Docker API or vulnerable container), there are a number of subsequent possible attack vectors. For instance, (Hertz 2016) describes a number of attacks for escaping privileged containers or exploiting unprivileged containers. (Avraami 2019) details how to abuse a specific CVE, CVE-2019-5736 to break out of Docker, (Sharma 2020) provides a walk-through for abusing one of the Linux Capabilities granted by default to Docker images to break out of a container. (Wist 2020) found that "As many as 82% of certified images [on DockerHub] contain at least either one high or critical vulnerability".

5. Methodology

5.1 Test Environment, Version Assumptions and Setup

We used CentOS as the Host OS for Docker. CentOS 7 was installed using the Net Installer CD, using mirror.centos.org/centos/7/os/x86_64 as the install source to ensure the latest release is obtained. As of 29 October 2020, the latest release version of Docker is 19.0.3. Docker was installed using the first set of steps presented on the Docker for CentOS install page 13. The VM was then powered down and a snapshot taken for quick reversion of state as needed.

5.2 Empirical Approach

There are a number of well-known issues with Docker containers which will be investigated in Section 6, Docker Security Analysis. These issues are composed of both weaknesses in the default security posture of a Docker deployment, as well as issues introduced through misconfigurations. We will investigate, either by practical demonstration or by reference, examples of these issues.

5.3 Theoretical Approach

Not all of the issues with Docker can be easily demonstrated for two reasons. Firstly, they either have not been discovered yet, or secondly, they have been patched out of the latest release of Docker, which this paper will use to establish a reasonable baseline of a Docker deployment. However, by studying past vulnerabilities and how they have exploited Docker, we can infer reasonable steps that may be taken to lockdown elements favoured for exploit and help mitigate the risk of future exploits.

6. Docker Security Analysis

This section will present some of the more common issues affecting Docker default deployments, however this does not claim to be a complete list of issues. Results will be analysed later to draft the proposed framework 'Zero Trust Container Architecture'. It should be noted, that rather than focus on specific attacks as some other papers have done, this paper will attempt to separate the components of a Docker deployment into 'trust zones', i.e., to consider the elements of a Docker deployment with respect to the trust, implicit or otherwise, that can be used to exploit the system, rather than to focus first on the attack and then the affected element. Decisions on what to investigate here were informed by the literature review.

6.1 Docker Service & RemoteAPI:

Misconfiguration of the Docker RemoteAPI/Service allowing public, unauthenticated access provide easy access for attackers to compromise the Docker deployment.

6.2 Users and the Docker Group:

Users are often the weakest element in any system. Docker is no exception and further aggravates this by treating all users as root allowing malicious, or negligent users to easily compromise a system.

6.3 Container Networking:

Docker's flat networking approach by default opens up containers to easy attack from other malicious containers. Several basic ICMP and TCP/IP attacks can be launched, that can be very difficult to detect inside a Docker network leading to silent and invasive compromise.

6.4 Container Resources:

Containers share the Host OS's available resources. By default, no resource usage limitations are placed on containers. This can be easily abused by malicious containers to perform Denial of Service, or potentially theoretical entropy attacks against legitimate containers.

6.5 Linux Capabilities:

Capabilities are a permissions subsystem in use by Linux and used with Docker to delegate permissions to containers. These can be abused to escape a container and compromise the Docker Service or the host directly.

6.6 Leaky Secrets:

Secrets, that is usernames and passwords, or API token, to access other application and third-party services, are often passed into containers as environmental variables. Should a container, or deployment be compromised, these secrets may lead to additional resources being compromised.

6.7 'Privileged' Mode:

Privileged mode grants a container all permissions on the Host OS. Should a privileged container be compromised, the Host OS, by default, should be considered compromised.

6.8 Docker vulnerabilities and Container Houdini:

As with any system, Docker suffers from vulnerabilities, both its own and those affecting related systems. These vulnerabilities have been identified as potential causes of exploit, and indeed have been used in real-world exploits as well.

6.9 Container Images:

Container images are point in time, static files that contain executable code to bootstrap an OS image, often containing libraries to support development, or entire applications. These images are uploaded to either private or public repositories for use by internal or external users. These images, by accident or indeed malicious design, may contain vulnerabilities or malware. The older the image, the more likely that several serious vulnerabilities exist.

7. Proposed Solution: Zero Trust Container Architecture (ZTCA)

In general, many of the issues identified can be described as the security reality of a default deployment of the Docker engine, or due to misconfigurations of Docker. That is, the implicit trust that users of Docker are delegating to the default install state of the Docker engine. This is why ZTA, an architectural framework that speaks specifically about trust in systems and how to avoid security issues arising from this trust, is suitable to mitigate these issues. Similar to (Rose 2019), we will create a framework by abstracting the principals of traditional ZTA, with a focus on challenging this implicit trust delegated to default installs of Docker. To be clear, there is nothing wrong with trust, or delegating trust within a system, however, this trust must be deliberate, explicit, and scoped appropriately. A key rule in ZTA is that every request, no matter where it is made from, should be secured, or verified. In traditional ZTA terminology this speaks to requests not only from 'north/south' sources being secured (from outside into a network) but also 'east/west' requests (within the same network) being secured. ZTA moves the boundary of trust from the network to the users and the endpoints (Rose 2019). If we refocus, or abstract, the boundary of ZTA to the Docker Service itself, this brings a completely different lens into what constitutes a request, as well as the meaning of securing said requests. For example, a user interacting with the Docker service becomes an untrusted communication. The Docker service or host OS communicating with a container (for instance passing secrets to a container) becomes an untrusted communication. The Docker service communicating with a container registry becomes an untrusted communication. Indeed, a similar argument for abstraction, albeit with a different focus is, made by (Rose 2019) "to take this one step further, the word "resource" can be substituted for "data" so that ZT and ZTA are about resource access and not just data access." By re-purposing this idea of a 'resource' in ZTA as the individual elements interacting with the Docker service, as well as considering the service itself, and combining this with a lens of trust, it is possible to create a set of 'maxims' for scoping, and analysing, the trust boundaries of a Docker deployment (see Figure 4).

This list of maxims allows for the design of a category/consideration matrix where elements of a Docker deployment, or 'Categories', are attributed a non-finite list of trust items, or 'Considerations'. 'Considerations' are suggested items to be considered when the related 'Category' is used with the framework, as such they are not mandatory, as they are not necessarily relevant to every scenario. Users of the framework may also add additional considerations to reflect their particular needs (e.g., alignment with another framework, standard or internal business rules).



Trust not the Host: The host running the Docker engine itself is a prime target for exploit. Docker does not exist in a vacuum and indeed the host it is running on has its own set of vulnerabilities (in particular kernel vulnerabilities as these increase the risk posed by compromised containers as they share the kernel with the Host) and security misconfigurations that allow it to be exploited and compromised independently of Docker. Securing the host is a crucial part of ensuring a Docker deployment remains secure.

.....



Trust not the User: By the terms of a default install of Docker, users granted access to the Docker group are effectively granted root access to the Host. This mandates much more consideration when granting users access to avoid unintentionally elevating permissions. Separately, users within a Docker container map to users on the host. By default, the lack of user namespacing means that running as root inside a container is effectively running as root on the Host (limited only by Container assigned CAPs on the root host).

.....



Trust not the Service: The Docker service runs with root permissions which can lead to a number of the compromises discussed previously in Section 6. Further, misconfigurations in presenting the Docker service publicly can lead to serious compromise with little to no effort on the part of the attacker.

.....



Trust not the Image: Docker images have been identified as a major concern with respect to trust. Both in terms of inherent vulnerabilities with images, or with outright malicious images. Further, a failure of process to frequent update images to the latest version could expose what is otherwise a secure system.

.....



Trust not the Container: By default, a Docker container itself presents a number of issues when it comes to implicit trust. We are trusting the lack of namespacing which allows container root to be restricted only by the CAPs applied to a container. We are trusting the Container not to abuse or overuse system resources. We are trusting a Container with secrets passed that may expose additional services to compromise. We are potentially trusting the container with privileged execution, grating it all power on the system. We are trusting it to share the kernel with the Host. All of these items must be properly reviewed and considered.

.....



Trust not the Network: By default, the flat network topology inside a single Host node allows for all Containers to communicate with each other, an implicit trust. Basic principles of ZTA dictate the need for securing east/west traffic, as well as north/south. As such this default network topology stands in violation of some of the fundamental rules of ZTA. Further, by default, there is no separation of concerns as with more traditional networks, which are often layered into distinct zones (LAN, DMZ, etc.) to limit communications and exposure, between services to only that which is necessary.

.....



Trust not the Silence: This describes the absence of logs, alerts and generally visibility, termed in this work simply as the 'Silence'. By default, Docker and the OS both provide some level of logging, however by default this is broadly opaque and decentralised. This does not meet the more comprehensive, auditable logging advised by ZTA. Docker logs are broadly concerned with the execution and state of a container, more specifically the 'STDOUT' and 'STDERR' logs, or with collecting metrics, no logging takes place with regards to network packets, traffic flow or inter-container communication. Without appropriate logging it is extremely difficult to understand what is actually taking place, to evaluate the situation and to ensure that trust has been delegated properly and is not abused.

Figure 4: ZTCA Maxims

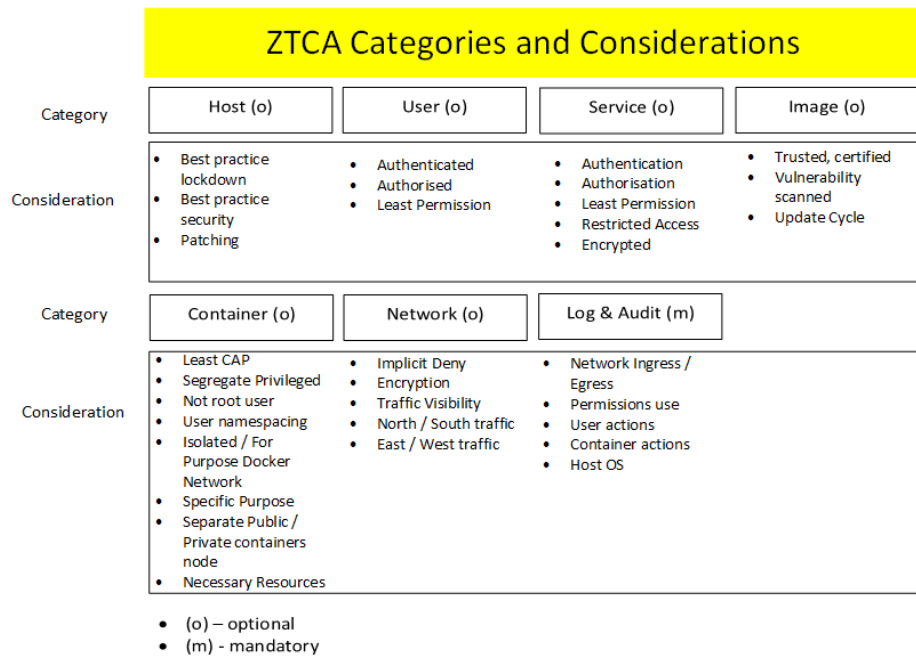


Figure 5: Categories and Considerations Matrix

The 'Categories and Considerations' matrix (Figure 5) serves as the input into the framework. Individual, or multiple, optional categories from the 'Categories and Considerations' matrix are passed to the framework for consideration. In order to simplify the use of the framework it is not advised to pass any more than 2 optional categories at a time. 'Log & Audit' category, which is mandatory, is to always be included.

7.1 ZTCA Framework

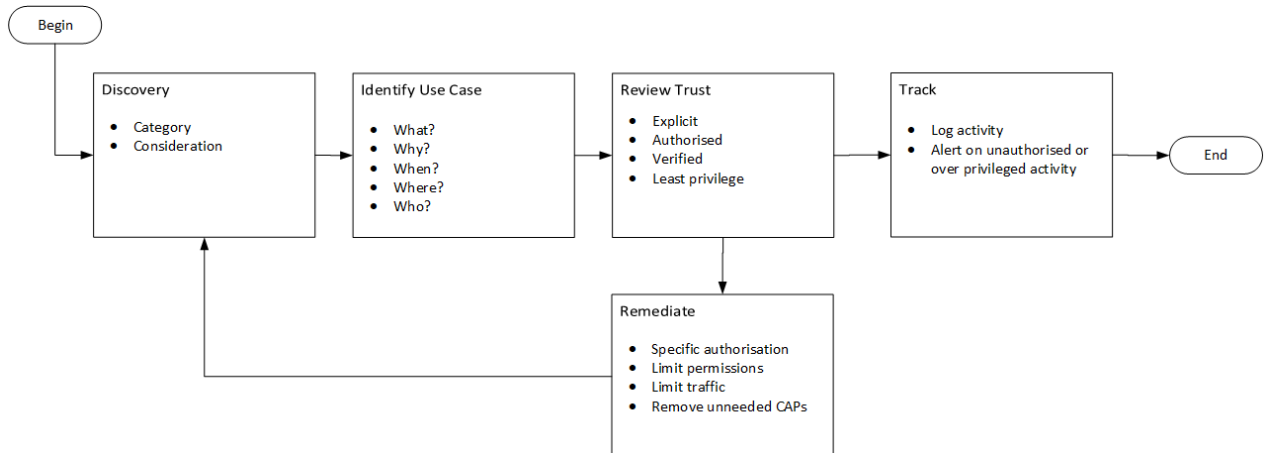


Figure 6a: ZTCA Framework

The framework (Figure 6a and 6b) will not attempt to be prescriptive, in that it will not specifically limit use cases (e.g. "never publicly expose the Docker Remote API") but will instead attempt to frame the thought process surrounding these use cases with respect to trust in such a manner as to ensure they are implemented in a secure fashion (e.g. "ensure the Docker RemoteAPI accepts only encrypted traffic, from only expected sources, from only authorised users"). As such, the framework does not output technical steps, as these may be specific to the users adopting the framework, for instance different firewall vendors mandate different technical steps. The framework is intended to be used in two ways. First, the framework can be used with a specific question, or use case in mind, for example 'grant a user access to the RemoteAPI', which the framework then can refine through a scoped lens of trust related to the categories passed, review the existing deployment, and propose remediation's to align the deployment with the refined use case. Secondly, the framework can be used to generate novel questions by passing category(s) without a specific question, or use case, in mind, and allowing users to consider their deployments through the lens presented by the framework and category(s) passed in order to generate novel questions. The framework operates by initially taking in category(s) which are used to

help scope and understand the question of trust being asked, then passing to subsequent functions which refine the exact use case in question, identify and define the appropriate trust required and contrast this with the existing trust in a deployment, before moving to a remediation step and finally the implementation of appropriate technical measurements to track the application of this trust. There are several decision gates included in the flow of the framework which can be used to iterate on sections which do not initially produce a refined output, or to review the process in light of remediation findings to ensure the framework has appropriately addressed all aspects of trust in question. The framework is designed to be used in conjunction with, not replace, other frameworks concerned with other elements of security (for instance frameworks for designing and securing a corporate network).

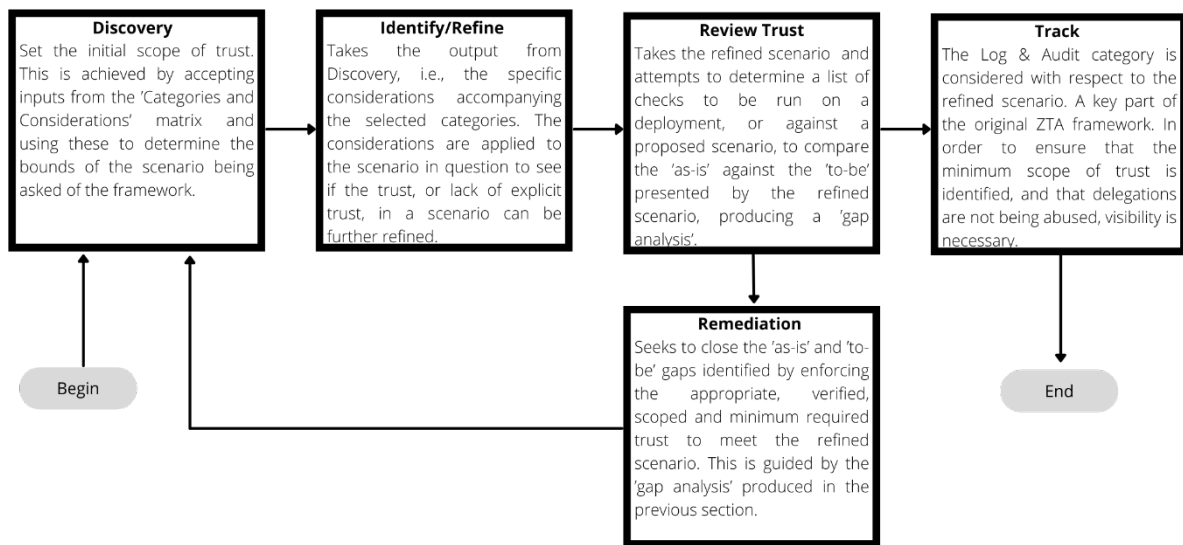


Figure 6b: ZTCA Framework Explained

8. Validation

Using the framework proposed, we analysed several scenarios derived from the issues identified in Section 6 and validated that the use of ZTCA leads to a successful mitigation, however it is not possible to address all attack scenarios demonstrated in Section 6, due to complexity and time required to implement and demonstrate mitigation of all in detail, however the following scenarios were investigated in detail:

- Test Scenario 1: Grant user access to Remote API (see Section 6.1)
- Test Scenario 2: Prevent Container Network Attack (see Section 6.3)
- Test Scenario 3: Review a Docker image (see Section 6.9)

Additionally, the following scenarios were walked through at a high level, that is the steps the framework suggests but without actually applying the mitigation:

- Test Scenario 4: Review an existing Container; a novel use case
- Test Scenario 5: Deploy a Privileged Container; as seen in Section 6.7 In beginning validation, it is worth noting that while there are various products on the market from a number of vendors to assist in securing aspects of Docker containers, or securing the network or Host OS, this paper will avoid providing implementation specifics for these products as this is outside the scope of this paper, but may reference concepts or utilise certain products as an example of how they can be applied to achieve compliance with suggestions of this framework.

This validation work is detailed in <https://bit.ly/3EPJXN0>, but due to space constraints, it cannot be presented in this paper.

9. Conclusion

This paper focussed on three research questions: (1) What issues currently affect a default deployment of Docker? (2) What additional common misconfigurations affect a Docker deployment? (3) Can ZTA, or an abstraction thereof be applied to help secure a Docker deployment? We have shown that Docker contains a

number of security concerns. From combination of both design issues and vulnerabilities, but also common misconfigurations that sacrifice security for convenience. Further, several security issues arise from the difference between a system espoused and a system in use. By distilling the principals of Zero Trust Architecture and leveraging these to create a ZTCA framework, this paper has proposed a method for analysing and determining the path to securing a Docker deployment. Validation has shown, when applying the framework to areas of concern as identified in Section 6, the proposed framework can be used to successfully analyse and recommend effective steps to ensure the security of a Docker deployment. The framework proposed, by virtue of the scope of possibilities, is complicated and assumes a level of proficiency with Docker as well as the ability to identify, analyse and implement additional appropriate security technologies. One of the main aims for future research should either be around a simplification of the security by design process, though this may rely on Docker Inc. developers themselves, or an abstraction layer for the Docker Engine which enforces the principals of security and least privilege by default.

References

- Avrahami, Y. (2019). Breaking out of docker via runc – CVE-2019-5736. <https://unit42.paloaltonetworks.com/breaking-docker-via-runc-explaining-cve-2019-5736/>. Accessed: 2022-01-06.
- Bettini, A. (2015). Vulnerability exploitation in docker container environments. <https://www.blackhat.com/docs/eu-15/materials/eu-15-Bettini-Vulnerability-Exploitation-In-DockerContainer-Environments-wp.pdf>.
- Bui, T. (2015). Analysis of docker security. <https://arxiv.org/pdf/1501.02967.pdf>. Accessed: 2022-01-06.
- de Weeve, C. and Andreou, M. (2020). Zero Trust Network Security Model in containerized environments. PhD thesis, University of Amsterdam.
- Graber, S. (2016). LXD 2.0: Introduction to LXD [1/12]. <https://stgraber.org/2016/03/11/lxd-2-0-introduction-to-lxd-112/>. Accessed: 2022-01-06.
- Hertz, J. (2016). Abusing privileged and unprivileged linux containers. Whitepaper, NCC Group, 48.
- Higashi, M., Kumar, G., and Chiodi, M. (2018). Lessons from the cryptojacking attack at tesla. <https://redlock.io/blog/cryptojacking-tesla>. Accessed: 2022-01-06.
- Kerman, A., Borchert, O., Rose, S., Division, Eileen, and Tan, A. (2020). Implementing a zero-trust architecture. Technical report, NIST
- Kindervag, J. (2010). No more chewy centers: Introducing the zero-trust model of information security. Forrester Research.
- Martin, A., Raponi, S., Combe, T., and Di Pietro, R. (2018). Docker ecosystem – vulnerability analysis. Nokia Bell Labs, 122
- Oliveira, A. (2019). Infected containers target docker via exposed APIs. https://www.trendmicro.com/en_us/research/19/e/infected-cryptocurrency-mining-containers-target-docker-hosts-with-exposed-apis-use-shodan-to-find-additional-victims.html. Accessed: 2022-01-06
- Quist, N. (2019). Misconfigured and exposed: Container services. <https://unit42.paloaltonetworks.com/misconfigured-and-exposed-container-services/>. Accessed: 2022-01-06.
- Rose, S., Borchert, O., Mitchell, S., and Connelly, S. (2019). Zero trust architecture. Technical report, National Institute of Standards and Technology.
- Simonovich, V. and Nakar, O. (2019). Hundreds of vulnerable docker hosts exploited by cryptocurrency miners. <https://www.imperva.com/blog/hundreds-of-vulnerable-docker-hosts-exploited-by-cryptocurrency-miners/>. Accessed: 2022-01-06.
- Sultan, S., Ahmad, I., and Dimitriou, T. (2019). Container security: Issues, challenges, and the road ahead. IEEE Access, 7:52976–52996
- Surantha, N. and Ivan, F. (2020). Secure kubernetes networking design based on zero trust model: A case study of financial service. Master's thesis, Bina Nusantara University, Jakarta, Indonesia.
- Yasrab, R. (2018). Mitigating docker security issues. <https://arxiv.org/ftp/arxiv/papers/1804/1804.05039.pdf>. Accessed: 2022-01-06.
- Wist, K., Helsem, M., and Gligoroski, D. (2020). Vulnerability analysis of 2500 docker hub images

APT cyber-attack modelling - building a general model

Martti Lehto

University of Jyväskylä, Finland

martti.j.lehto@jyu.fi

Abstract: The global community continues to experience an increase in the scale, sophistication, and successful perpetration of cyber-attacks. As the quantity and value of electronic information have increased, so too have the efforts of criminals and other malicious actors who have embraced the Internet as a more anonymous, convenient, and profitable way of carrying out their activities. The systems are attacked more and more by single or multiple hacktivists, state sponsored hackers, cyber criminals, cyber terrorists, cyber spies, or cyber warfare warfighters. The cyber security approach requires a balance of cyber threat intelligence, real time cyber-attack detection and especially the cyber early warning ability. Threats in cyberspace are difficult to define, as it is hard to identify the source of attacks and the motives that drive them, or even to foresee the course of an attack as it unfolds. The identification of cyber threats is further complicated by the difficulty in defining the boundaries between national, international, public, and private interests. Because threats in cyberspace are global in nature and involve rapid technological developments, the struggle to respond them is ever-changing and increasingly complicated. Cyber-attack models describe the structure of an attack in different phases. They provide a means to conceptualize the different aspects and elements of an attack. However, it is important to understand that not all attacks must complete all phases to be successful, and the objective of the attack defines the structure of the attack. Different actors have built different cyber-attack models. Modeling is used to understand the different goals of cyber attackers. Attack models are based on attack targets and attack objectives. This paper analyzes different APT cyber-attack models and presents a general cyber-attack model.

Keywords: cyber security, cyber-attack model, ATP attack

1. Introduction

Attack models describe the structure of an attack in different phases. They provide a means to conceptualize the different aspects of an attack. However, it is important to understand that not all attacks must complete all phases to be successful. In fact, many attacks iterate recursively through the phases of an attack model.

The models have been developed based on the motives and objectives of various cyber-attacks. Cyber-attacks are often divided four categories based on the attack objective and target: APT-attack, Cyber-Physical Attack against Critical Infrastructure, Data breach attack, and Military Cyberspace Operations.

The objective of this paper is to define the general cyber-attack model in an APT-threat context. Section 2 explains into ATP attack essence and in Section 3 is describing existing ATP attack models. Section 4 addresses general cyber-attack modeling., followed by a discussion of the relevance of the modelling of the cyber-attacks.

2. APT attack essence

An Advanced Persistent Threat (APT) is a stealthy computer network threat actor, typically a nation state or state-sponsored group, which gains unauthorized access to a computer network and remains undetected for an extended period. In recent times, the term may also refer to non-state sponsored groups conducting large-scale targeted intrusions for specific goals. APT is a prolonged and targeted cyberattack in which an intruder gains access to a network and remains undetected for an extended period of time. Normally APT attacks are initiated to steal data rather than cause damage to the target organization's network (Fire Eye, 2021).

The goal of most APT attacks is to achieve and maintain ongoing access to the targeted network rather than to get in and out as quickly as possible. The attackers use a great deal of effort and resources to carry out APT attacks, and they typically select high-value targets. The targets of APT attack include large organizational networks that contain valuable secret information. APT attacks can be defined as action aimed at acquiring secret information (sensitive, proprietary, or classified) from competitors, groups, governments, and adversaries for political, military, or financial gain by using special methods on the Internet, in networks, programs or computers (Liaropoulos 2010); Segal, 2019).

Executing an APT attack requires a higher degree of customization and sophistication than a traditional attack. APT groups are a nation state or state-sponsored who have advanced devices and are already financially

covered. Several effective APT groups have been identified from Russia and China. Advanced APT attack operations will remain secret for several years. Table 1 illustrates this detection problem with a few examples.

Table 1: Apt attack operations

APT operation	Targets	Year of detection	In operational use since
Operation Aura	Dozens of IT organizations (Google among others) specially in US.	2010	2009
Operation Shady RAT	More than 70 companies and government entities around the world.	2011	2007
Duqu	Information from industrial control systems.	2011	2007
skyWiper/Flame	The Middle East governmental organizations, educational institutions, and individuals.	2012	2008
NetTraveler	More than 350 organizations in 40 countries	2013	2004
Regin	Several organizations at least 14 countries like small businesses and telecom companies.	2014	2008, first sample 2003
Remsec (Project Sauron)	More than 30 organizations in Russia, Iran, China: government, scientific research centers, military, telecommunication providers and finance.	2016	2011

3. ATP-attack models

The APTs are flexible and sophisticated. Skilled and determined cyber attacker can use multiple vectors and entry points to navigate around defenses, breach the network in minutes and evade detection for months or even years. The process by which sophisticated cyber-attacks are conducted can be described as a lifecycle. Due to the diversity of APT attacks, several models have been developed:

1. MITRE ATT&CK
2. Mandiant Attack Lifecycle Model
3. LM Cyber Kill Chain, CKC
4. Unified Kill Chain, UKC
5. Hybrid Cyber Kill Chain, HCKC

3.1 MITRE ATT&CK

MITRE ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations. The ATT&CK knowledge base is used as a foundation for the development of specific threat models and methodologies in the private sector, in government, and in the cybersecurity product and service community. The aim of the framework is to improve post-compromise detection of adversaries in enterprises by illustrating the actions an attacker may have taken (MITRE, 2021).

The model has two main phases. Pre-phase covering preparatory techniques: reconnaissance and weaponization. Adversary pre-phase activities are largely executed outside the organizations' field of view, making them more difficult to detect. In the second phase, the attack phase itself is modeled. The framework consists of 11 tactics and one objective of the mission (MITRE, 2021).

1. Initial Access: The adversary is trying to get into your network.
2. Execution: The adversary is trying to run malicious code.
3. Persistence: The adversary is trying to maintain their foothold.
4. Privilege Escalation: The adversary is trying to gain higher-level permissions.
5. Defense Evasion: The adversary is trying to avoid being detected.
6. Credential Access: The adversary is trying to steal account names and passwords.
7. Discovery: The adversary is trying to figure out your environment.
8. Lateral Movement: The adversary is trying to move through your environment.
9. Collection: The adversary is trying to gather data of interest to their goal.
10. Command and Control: The adversary is trying to communicate with compromised systems to control them.
11. Exfiltration: The adversary is trying to steal data.
12. Impact: The adversary is trying to manipulate, interrupt, or destroy your systems and data.

MITRE ATT&CK was created in 2013 to improve post-compromise detection of threats through telemetry sensing and behavioral analysis. ATT&CK is also used as a tool to categorize adversary behavior (McAfee, 2021).

3.2 Mandiant Attack Lifecycle Model

APT attacks fit into a cyclic pattern of activities that the framework of Mandiant's Attack Lifecycle model describes. The stages between "Establish Foothold" and "Complete Mission" do not have to occur in this order every time. In fact, once established within a network, APT groups will continually repeat the cycle of conducting reconnaissance, identifying data of interest, moving laterally to access that data, and "completing mission" by stealing the data. This will generally continue indefinitely until they are removed entirely from the network (IACP, 2021; Parrend et al, 2018).

The model consists of eight stages:

1. Initial Reconnaissance: The attacker conducts research on a target. The attacker identifies targets (both systems and people) and determines his attack methodology.
2. Initial Compromise: The attacker successfully executes malicious code on one or more systems. This most likely occurs through social engineering.
3. Establish Foothold: The attacker ensures he maintains continued control over a recently compromised system. Typically, the attacker establishes a foothold by installing a persistent backdoor or downloading additional utilities or malware to the victim system.
4. Escalate Privileges: The attacker obtains greater access to systems and data. Attackers often escalate their privileges through using vulnerabilities in software and/or hardware.
5. Internal Reconnaissance: The attacker explores the victim's environment to gain a better understanding of the environment, the roles, and responsibilities of key individuals, and to determine where an organization stores information of interest.
6. Move Laterally: The attacker uses his access to move from system to system within the compromised environment.
7. Maintain Presence: The attacker ensures continued access to the environment. Common methods of maintaining a presence include installing multiple variants of malware backdoors or by gaining access to remote access services.
8. Complete Mission: The attacker accomplishes the goal. Once the mission has been completed, most targeted attackers do not leave the environment, but maintain access in case a new mission is directed.

3.3 LM Cyber Kill Chain (CKC)

The Cyber Kill Chain framework is developed by Lockheed Martin for identification and prevention of cyber intrusions activity. The model identifies what the adversaries must complete to achieve their objective. The seven steps of the Cyber Kill Chain enhance visibility into an attack and enrich an analyst's understanding of an adversary's tactics, techniques, and procedures (LM, 2021).

1. Reconnaissance: Identify the Targets: The adversaries are in the planning phase of their operation. They conduct research to understand which targets will enable them to meet their objectives.
2. Weaponization: Prepare the Operation: The adversaries are in the preparation and staging phase of their operation. A "weaponizer" couples malware and exploit into a deliverable payload.
3. Delivery: Launch the Operation: The adversaries convey the malware to the target. They have launched their operation: Adversary controlled delivery or adversary released delivery.
4. Exploitation: Gain Access to Victim: The adversaries must exploit a vulnerability to gain access. Zero-day vulnerabilities are exploited in this step.
5. Installation: Establish Beachhead at the Victim: Typically, the adversaries install a persistent backdoor or implant in the victim environment to maintain access for an extended period.
6. Command & Control (C²): Remotely Control the Implants: Malware opens a command channel to enable the adversary to control the operation.
7. Actions on Objectives: Achieve the Mission's Goal: The attacker accomplishes the mission's goal, like collect and exfiltrate data, destroy systems, overwrite, or corrupt data.

3.4 Unified Kill Chain, UKC

The Unified Kill Chain (UKC) was proposed in 2017 by Paul Pols. Combining elements from both the Cyber Kill Chain and ATT&CK (along with several other models), it divides an attack into 18 separate stages. Its aim is to expand the scope of the Cyber Kill Chain and to represent an improvement over the time-agnostic nature of ATT&CK. UKC model is developed that focuses on the tactics that form the consecutive phases of cyber-attacks. The resulting UKC is a meta model that supports the development of end-to-end attack specific kill chains and actor specific kill chains, that can subsequently be analyzed, compared, and defended against (Pols, 2017; Pols, 2019).

The 18 steps are broadly grouped into three overarching areas. First, the *Initial Foothold* must be gained. This phase covers the steps taken to compromise a single system within a target and includes steps 1-8. In *Network Propagation* the attacker attempts to pivot from their compromised system to other, more important parts of the target's environment. It comprises the steps 9-14. The *Action on Objectives* describes a situation in which the attacker, having found the goal, gains privileged access to a part of critical infrastructure, and performs initially planned task. This could be the exfiltration of sensitive data or its manipulation. It comprises the steps 15-18. (Pols, 2017; Pols, 2019)

The 18 stages of the Unified Kill Chain are (Pols, 2019):

1. Reconnaissance: Researching, identifying, and selecting targets using active or passive reconnaissance.
2. Weaponization: Preparatory activities aimed at setting up the infrastructure required for the attack.
3. Delivery: Techniques resulting in the transmission of a weaponized object to the targeted environment.
4. Social Engineering: Techniques aimed at the manipulation of people to perform unsafe actions.
5. Exploitation: Techniques to exploit vulnerabilities in systems that may, amongst others, result in code execution.
6. Persistence: Any access, action or change to a system that gives an attacker persistent presence on the system.
7. Defense Evasion: Techniques an attacker may specifically use for evading detection or avoiding other defenses.
8. Command & Control: Techniques that allow attackers to communicate with controlled systems within a target network.
9. Pivoting: Tunneling traffic through a controlled system to other systems that are not directly accessible.
10. Discovery: Techniques that allow an attacker to gain knowledge about a system and its network environment.
11. Privilege Escalation: The result of techniques that provide an attacker with higher permissions on a system or network.
12. Execution: Techniques that result in execution of attacker-controlled code on a local or remote system.
13. Credential Access: Techniques resulting in the access of, or control over, system, service or domain credentials.
14. Lateral Movement: Techniques that enable an adversary to horizontally access and control other remote systems.
15. Collection: Techniques used to identify and gather data from a target network prior to exfiltration.
16. Exfiltration: Techniques that result or aid in an attacker removing data from a target network.
17. Impact: Techniques aimed at manipulating, interrupting or destroying the target system or data.
18. Objectives: Socio-technical objectives of an attack that are intended to achieve a strategic goal.

The main advantage of the way the Unified Kill Chain is laid out is in the use of repetition. Whereas the Cyber Kill Chain suggests a step 1, 2, 3 approach to attacking, the Initial Foothold, Network Propagation and Action on Objectives stages of UKC are presented as loops that may, eventually, move on to the next stage. This better captures the actual behavior of an attacker, who is likely to spend a lot of time attempting to gain that initial foothold and comparatively less actually doing what he comes to do (Pols, 2019).

3.5 Hybrid Cyber Kill Chain, HCKC

Zeng and Germanos (2019) constructed a Hybrid Cyber Kill Chain (HCKC) framework. Although LM Cyber Kill Chain has been widely adopted by organizations since it has been published, it has been criticized for its weakness in handling insider threats, because it is too focused on the perimeter and malware prevention. For example, shows that the stages of CKC do not capture all possible attack vectors and are not expressive enough to model internal actions within the target network. Consequently, based on these facts, improvements, and amendments to CKC have been proposed. Some of these proposed extensions try to address the insider's perspective (Zeng and Germanos, 2019).

HCKC framework aims to provide a holistic approach by considering the factors, like the persistence and the lateral movement of the attacker, the defence actions of the defender, and the strength level both attacker and defender. The third factor is embedded in each stage of HCKC after the weaponization. That is because the two first stages are not under the control of the defender. According to this rule, the attacker's strength level must be greater than the one of the defenders to proceed to the next stage in the kill chain (Zeng and Germanos, 2019).

Stage 1 - Reconnaissance: The attackers would gather as much information as possible about target organization. For example: names of the employers, their positions, email addresses and IP addresses.

Stage 2 - Weaponization: The attackers develop exploitation and malware. The attackers might create websites that contain malware and develop malicious software for a specific platform or purpose, which designed according to the vulnerabilities discovered during reconnaissance.

Stage 3 - Delivery: The attackers transmit the malicious code to the target information system. The attacker might use a spear-phishing attack targeting an internal employee of the organization, social engineering, or some vulnerability of the system to deliver the malware.

Stage 4 - Persistence: This stage denotes the attacker's persistence to accomplish an action.

Stage 5 - Exploitation: The attacker utilizing the discovered vulnerabilities, is executing the malicious code on the target network using remote or local mechanisms. The aim is to gain admin access to the target's information system.

Stage 6 - Installation: After the exploitation, the malicious code will install itself onto the target's information system. Then, the malicious code will start downloading additional software when there is available network access.

Stage 7 - Command & Control: The attacker has set up the management and communication code onto the target network. Now, the attacker can fully manage the malicious code and move further into the network, and exfiltrate data.

Stage 8 - Lateral Movement: This stage expresses the move of the attacker to bigger targets using the target network after some of its systems has been compromised.

Stage 9 - Action on Objectives: The actions and objectives of the attacks are dependent on their specific mission (e.g., exfiltrate or conduct disruption or destruction of the data).

Zeng and Germanos (2019) states that HCKC is useful for post-incident analysis of attacks. A formal model of the HCKC can also provide security metrics that can be utilized by system architects to make trade-off decisions involving system security. There is a growing need for efficient and timely incident response. In many cases, the period between detection and response can be months long and can allow adversaries to attain their goals (Zeng and Germanos, 2019).

4. General cyber-attack modelling

Cyber-attack models provide us with a means of decomposing an attack into discrete phases. These in turn can be used to conduct post-intrusion analysis to better predict and avoid future attacks. By understanding the tactics, techniques, and procedures (TTPs) of an attacker, we learn how threat actors operate and have the means to evaluate our defensive posture and develop strategic courses of action to eliminate gaps.

APT cyber-attack models presented above have common elements. The main differences in cyber-attack patterns are due to the level of detail of the attack operation. Some models include only the main components of the attack, while others describe the attack components accurately.

Based on the different ATP attack models, the general model has been developed. Other models do not present an early attack phase, i.e., strategic decision-making, that precedes the actual preparation for an APT attack. Cyber-attack can thus be divided into three main stages: early-attack, pre-attack, and the actual attack operation. Most models recognize pre-attack and attack phases. While most models recognize pre-attack and attack phases, other attack models do not describe very clear the end state phase, i.e., termination of the attack, which addresses the issues such as how the attacker completes the mission and how he covers his tracks. Different cyber-attack models use different terms to describe different stages of an attack.

In the early attack phase the attacker will made the strategic decision of the target and attack objectives. Strategic decisions involve a political problem of reconciling divergent interests as well as a technical problem of attempting to calculate the best decision given several parameters (Child, Elbanna and Rodrigues, 2010). Figure 1 illustrates the general cyber-attack model.

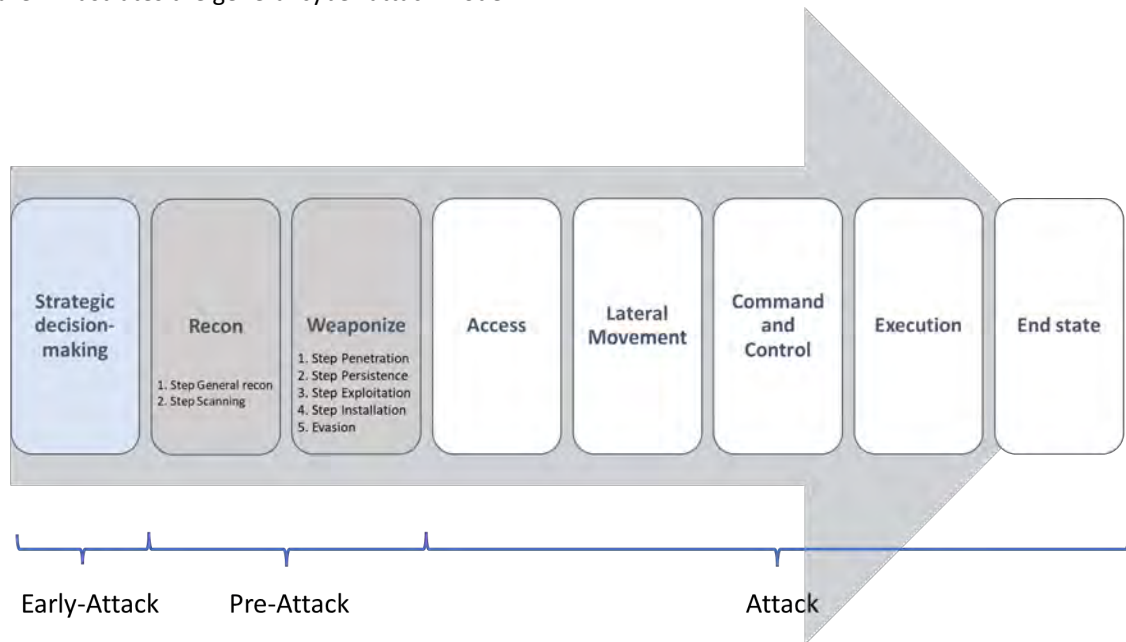


Figure 1: General Cyber-attack model

The table 2 illustrates the phases and steps of the general APT cyber-attack model.

Table 2: General APT cyber-attack model

General APT cyber-attack model	
Early-Attack phase	
Strategic decision-making	In state-sponsored APT attacks, a decision is made at the strategic level on the target of the attack and the objectives of the attack, and how it relates to other strategic objectives.
Pre-Attack phase	
Reconnaissance - target identification - target location - target characteristics	1st step: General reconnaissance: attack planners create the situational awareness from the target organization. They conduct research to understand which targets will enable them to meet their objectives. In reconnaissance phase the attacker want to gather as much information as possible about target organization. 2nd step: Scanning: attacker will be closely examining the system for potential vulnerabilities. The attacker will be scanning for further detailed information from applications, and more specific information from the operating system. The attacker focuses on finding an exposed application that might be particularly interesting in terms of entry.
Weaponize - operation preparing and malware production	In weaponization phase the attackers develop exploitation and malware. The attackers might create websites that contain malware or develop malicious software which designed according to the vulnerabilities discovered during reconnaissance. A weapon designer compiles malware and exploit into a deliverable payload (in-house production). Weapons are also available on the Dark web. For the file-based exploits the designer develops a dropper to hide the real payload. Also, the designer selects backdoor implant and appropriate command and control infrastructure for operation.
Attack phase	
Access - entry to target network - penetration - persistence - exploitation	The access phase represents the methods that attacker uses to penetrate a target organization's network and take it over to execute the operation. The attacker seeks to achieve a main admin authority. This phase is divided five steps. 1st step: Penetration: Attacker frequently targets individual users within a victim environment. As such, the most observed method of initial compromise is spear phishing. The attacker might

General APT cyber-attack model	
- installation - evasion	use a spear-phishing attack targeting an internal employee of the organization, social engineering, or some vulnerability of the system to deliver the malware. 2nd step: Persistence: the attacker transmits the malicious code to the target information system and establishes a foothold ensures that APT attack groups can access and control one or more computers within the victim organization from outside the network. Typically, the attacker installs a persistent backdoor or implant in the victim environment to maintain access for an extended period. 3rd step: Exploitation: the attacker utilizes the discovered vulnerabilities to gain access, executes the malicious code on the target network. The aim is to get administrator rights to the target's information system. Vulnerabilities be found from people, technologies, or processes. The most effective are the zero-day vulnerabilities. 4th step: Installation: As soon as the exploitation is successful, the malicious code will install itself onto the target's information system. The attacker can maintain continuity of access to the target system by a remote access trojan or creating a backdoor to penetrate further into the target network. 5th step: Evasion: The attacker is trying to avoid being detected. The attacker tries to stay hidden by carrying out a stealth attack. State-Sponsored Attackers appear to be successful in this regard, as an average APT malware is found after five years of attack.
Lateral Movement - surveillance - environment expanding	In the lateral movement phase the attacker is trying to figure out ICT environment. Also, the attacker tries enlarging attack surface using the target network after some of its systems has been compromised. The attacker tries escalating privileges involves acquiring items that will allow access to more resources within the victim environment.
Command and Control - attack management - maintain presence	In the command and control (C²) phase the attacker will set up the management and communication code onto the target network. Then the attacker can manage the malware and move further into the network and execute the operation. In this phase, the attacker takes actions to ensure continued control over key systems in the network environment from outside of the network.
Execution - data collection - data exfiltration - data manipulation	In the execution phase the attacker executes the mission = action on objectives. The actions and objectives of the attacks are dependent on their specific mission: data collection and exfiltration, data manipulation (corruption) or data overwriting. The main goal of APT attack is to steal data, including intellectual property, business contracts or negotiations, secret policy and military papers and information etc. Today, data manipulation is considered one of the most dangerous attacks.
End state - termination of the attack	When the attack objectives are accomplished, the attacker disappears unnoticed. The attacker tries to clean-up all traces from the "visiting" in the ICT system. "Mission accomplished in stealth." If the APT attack involved a silent data exfiltration which was not detected, attackers will remain inside the network and wait for additional attack opportunities. Over time they may collect additional sensitive data and repeat the process. They will also aim to create backdoors that are difficult to detect, so even if they are caught, they can regain access to the system in the future.

Different cyber-attack models use different terms to describe different stages of an attack. At a high level of abstraction, different cyber-attack models contain almost identical steps. But there are also differences in definitions and descriptions of steps. Table 3 shows five APT attack models combined in a general model.

Table 3: APT-attack models versus general model

General Modell	MITRE ATT&CK	Mandiant Attack Lifecycle Model	LM Cyber Kill Chain, CKC	Unified Kill Chain, UKC	Hybrid Cyber Kill Chain, HCKC
Early Attack					
Strategic decision-making					
Pre-Attack					
Reconnaissance - target identification - target location - target characteristics	1. Reconnaissance	1. Initial reconnaissance	1. Reconnaissance	1. Reconnaissance	1. Reconnaissance
Weaponize - operation preparing and malwares	2. Weaponization		2. Weaponization	2. Weaponization	2. Weaponization
Attack					
Access - entry to target network - penetration - persistence - exploitation - installation - evasion	3. Initial access 4. Execution 5. Persistence 6. Privilege escalation 7. Defense evasion 8. Credential access 9. Discovery	2. Initial compromise 3. Establish foothold 4. Escalate privileges 5. Internal reconnaissance	3. Delivery 4. Exploitation 5. Installation	3. Delivery 4. Social engineering 5. Exploitation 6. Persistence 7. Defence evasion 8. Pivoting 9. Privilege escalation 10. Credential access 11. Discovery	3. Delivery 4. Persistence 5. Exploitation 6. Installation
Lateral Movement - surveillance - environment expanding	10. Lateral movement	6. Lateral movement		12. Lateral Movement	7. Lateral movement
Command and Control - attack management - maintain presence	11. Command and Control	7. Maintain presence	6. Command & Control	13. Command & Control	8. Command & Control
Execution - data collection - data exfiltration, - data manipulation	12. Collection 13. Exfiltration 14. Impact	8. Complete mission	7. Actions on objectives	14. Execution 15. Collection 16. Exfiltration 17. Target manipulation 18. Objectives	9. Action on objectives
End state - termination of the attack					

5. Conclusion and discussion

There are a wide range of internet threats and attacks from virus propagation and worms, to distributed denial of service (DDoS) attacks and data theft and manipulation and also critical infrastructure paralysis. So far, many proactive techniques have been proposed to deal with these threats. All these techniques pursue the same goal, preventing attackers from reaching their objectives.

There is great interest in developing proactive methods of cyber security, in which future attack strategies are anticipated and cyber-attack models are used. The APT attack model is crucial for cyber early warning systems.

Early warning is a proactive approach against security attacks and threats. These systems are a complement to intrusion detection and prevention systems where the main goal of them is the early detection of the potential behavior of a system, evaluating the scope of malicious behaviors, and finally applying a suitable response against any kind of detectable security event. APT attack modelling makes it possible to build efficient early warning system which collects, correlates the heterogeneous logs from various sensors and provides timely and effective information to avoid or reduce potential risks and prepare effective response with future behavior prediction. Modern cyber-attack warning should concentrate on the identification of indicators as early as possible to augment resilience.

Also, APT attack modelling helps to develop Cyber Threat Intelligence (CTI). It refers to a dynamic, adaptive technology that leverages large-scale threat history data to proactively block and remediate future malicious attacks on a network. Cyber threat intelligence recognizes indicators of attacks as they progress and essentially puts these pieces together with shared knowledge about attack methods and processes (Shackleford, 2015).

The general APT attack model helps in different levels of the CTI. Strategic level cyber threat intelligence informs the most senior decision-makers, in the operational level CTI is aimed at making day-to-day decisions and in tactical level CTI is focused on units in need of instantaneous information. The model can be applied to the needs of each CTI level.

References

- Child J., Elbanna S., and Rodrigues S. (2010) *The Political Aspects of Strategic Decision Making*, in the Handbook of Decision Making, Paul C. Nutt and David Wilson (eds.), Chichester: Wiley.
- Fire Eye. (2021) *Anatomy of Advanced Persistent Threat*, <https://www.fireeye.com/current-threats/anatomy-of-a-cyber-attack.html>, retrieved 29.9.2021.
- IACP. (2021) *Cyber Attack Lifecycle*, on-line information, <https://www.iacpcybercenter.org/resource-center/what-is-cyber-crime/cyber-attack-lifecycle/>, retrieved 17.10.2021.
- Liaropoulos A. (2010), *War and ethics in cyberspace: Cyber-conflict and just war theory*, in J. Demergis (ed.), Proceedings of the 9th European Conference on Information Warfare and Security, Greece.
- LM. (2021) *The Cyber Kill Chain®*, on-line information, <https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html>, retrieved 17.10.2021.
- McAfee. (2021) What is the MITRE ATT&CK framework? on-line information, <https://www.mcafee.com/enterprise/en-us/security-awareness/cybersecurity/what-is-mitre-attack-framework.html>, retrieved 20.10.2021.
- MITRE. (2021) *MITRE ATT&CK®*, on-line information, <https://attack.mitre.org/>, retrieved 19.10.2021.
- Parrend, P. & Navarro, J. & Guigou, F. & Deruyver, A. & Collet, P. (2018). *Foundations and Applications of Artificial Intelligence for Zero-day and Multi-Step Attack Detection*. EURASIP Journal on Information Security.
- Pols P. (2017) *The Unified Kill Chain Designing a Unified Kill Chain for analyzing, comparing, and defending against cyber-attacks*, Cyber Security Academy (CSA), December 7, 2017.
- Pols P. (2019) *The Unified Kill Chain – Raising Resilience against Advanced Cyber Attacks*, White Paper, <https://www.unifiedkillchain.com/assets/The-Unified-Kill-Chain.pdf>, retrieved 2.10.2021.
- Segal E. (2019) *APT Attacks 101: What They Are and How They Work*, internet blog, Sep 27, 2019, https://medium.com/@eddie_47682/apt-attacks-101-what-they-are-and-how-they-work-393c09f55eae, retrieved 10.10.2021.
- Shackleford D. (2015) *Who's Using Cyberthreat Intelligence and How?* SANS, February 2015
- Zeng W. and Germanos V. (2019) *Modelling Hybrid Cyber Kill Chain*, Cyber Technology Institute, School of Computer Science and Informatics De Montfort University, Leicester.

Design and Evaluation of a Cyber Protection Team Planner Work Aid

Kristen Liggett, Arielle Stephenson, Meghan Strang and Geoffrey Dobson

Air Force Research Laboratory, WPAFB, OH USA

Kristen.Liggett@us.af.mil

Arielle.Stephenson.ctr@us.af.mil

Meghan.Strang.1@us.af.mil

Geoffrey.dobson.1@us.af.mil

Abstract: Cyber Protection Teams (CPTs) are responsible for providing mission assurance by ensuring that mission-relevant cyber terrain is secure and protected for operations. In order to do this effectively, CPT planners must correlate and integrate information regarding customers' missions with information about the systems and networks required to perform those missions. To support this challenge, a CPT Planner work aid was designed and evaluated to assess its effectiveness. The evaluation compared the work aid with current CPT planning methods. Ten participants with varying planning expertise performed four planning tasks (managing approvals, assigning resources, managing accesses, and scheduling meetings). To test the efficacy of the proposed work aid, participants were tasked with planning a fictional mission utilizing both the traditional planning method and the work aid method. Data was collected from objective metrics to include time and task performance score, as well as subjective perception of workload. In addition, participants were asked a series of situation awareness (SA) questions after each task to measure their understanding of the mission-relevant information. Score and time to answer SA questions were also recorded. While CPT planning experience did not have a significant effect on performance, results revealed significant improvements in task time, SA question response time, and perceived workload when using the work aid versus the baseline condition.

Keywords: cyber protection teams, work aid design, situation awareness metrics

1. Background

The cyber domain is one of incessant progress and growing emphasis in warfare operations. Within this dynamic field, cyber security and cyberspace operations sit in a state of unrest. Cyberspace is defined by Air Force Doctrine Annex 3-12 Cyberspace Operations as a "global domain within the information environment consisting of the interdependent network of information technology infrastructures, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers" (United States, 2011). Cyberspace operations consist of military activities to identify, degrade and/or deceive threat actors in a virtual space, while cyber security functions to ensure network operations are trustworthy (Trent et al, 2016). As technology and computers continue to play a prominent role in everyday activities across the world, the Department of Defense (DoD) has recognized the potential threats that accompany this surge in cyber usage. In order to defend and protect against these inherent risks, U.S. Cyber Command was established in 2009 and further expanded into multiple task groups including Cyber Mission Teams and a Cyber National Mission Force in 2012 (Lathrop, Trent, and Hoffman, 2016; U.S. Army Cyber Command, 2020).

The DoD's establishment of cyber-specific tasks groups has been a critical advancement in the endeavour to maintain cyberspace dominance. However, these unique groups have emerging capabilities and therefore new requirements that must be assessed. For example, operators in the field require certain skillsets and tools in order to properly assess situations and determine appropriate courses of action. Operators have reported that even when tools are accessible, they are often non-functional, are not fully understood by the operator(s), or are becoming outdated (Anderson, 2018). Additional unique limitations arise from the design space for these tools. For example, cyberspace tools are often built by and for networking experts who have full privileges and access within the network while they are designing the tool, and these conditions rarely exit for teams in an operational setting (Trent et al, 2016). Additionally, a qualified operator may have accessibility to the necessary tools, but obtaining the required permissions to utilize them cause a major setback in operations. As stated in *Cyber Network Mission Dependencies*, "Because all DoD missions depend on cyber infrastructure, failure to secure network assets and assure the capabilities they enable will pose a fundamental risk to any defense mission," (Schulz, Kotsin, and Gibson, 2015). With rapidly evolving technological advancements across the globe, it is critical that operators are equipped with useful and usable tools in order to remain vigilant.

The Cyber Mission Force is constructed of various types of teams, each with respective assignments that contribute to the overall mission. Among the support teams are Cyber Protection Teams (CPT), which serve to defend information networks, prepare operational cyber forces, and protect high priority missions (U.S.

Congressional Research Service, 2020). CPTs must provide expert skills and core competencies in order to mitigate threats and assure mission readiness. The four main mission types conducted by CPTs include network assessments, incident response, threat assessment, and active defense (Lathrop, Trent, and Hoffman, 2016). As of 2019, CPTs lack a planning tool and instead rely on a shared information repository (Pomerleau, 2019). This contains a combination of spreadsheets, typed documents, and hand-written visuals which provide the necessary information for planning missions. While this type of information allows the user access to available resources, the fact that they are all in different formats (some are hard copy only) and the information is not integrated results in increased planning time and cognitive load. Therefore, the objective of this study was to design and evaluate a CPT planning tool in order to increase the efficiency and effectiveness in which planners can execute their tasks.

2. Design process

The CPT Planner work aid was designed using the Human Factors Design Process. This process consists of an analysis phase and an iterative design phase. This section will describe these phases along with a description of the work domain analysis results.

2.1 Analysis phase

During the analysis phase, the work domain was analyzed by conducting literature reviews, participating in relevant workshops, attending demonstrations of relevant software, conducting targeted topic area discussions, participating in cyber exercises, and visiting various operational sites. Knowledge acquisition (KA) activities were conducted with a number of CPT planners and subject matter experts (SMEs) in the area of mission planning. Additionally, research was conducted at United States Transportation Command (USTRANSCOM) and firsthand observations were made of CPTs being tasked and responding to a mission. The design team also received artifacts from CPT planners that gave insight into the types and format of information required to perform their tasks, as well as how they administered or completed said tasks using current methodologies and practices.

2.2 CPT planner workflow

CPT planning is initiated with the receipt of a task order to support a particular mission. The CPT planner reviews the information provided on the task order and then determines what additional information is needed. This information must be requested from mission owners, network owners, and system owners, and is currently done via email to the various points of contact (POCs). Typically, POCs are identified in the task order, however, if they are not, the CPT planner must track them down. While waiting for information requests to be filled, the CPT planner reviews CPT actions that are pre-approved for the mission and listed on the task order. The planner determines if additional approvals for action on the mission relevant cyber terrain are required and will request additional approvals if necessary. The CPT planner cannot begin their mission until the appropriate approvals have been received from the POC. Following approvals, the planner must determine further resources that are required to conduct the mission. This includes determining which members are qualified to utilize the tools needed to perform necessary tasks, along with analyzing those members' availability. If certain tasks require various network or building access, these must also be requested based on task assignments. The final planning step includes scheduling a variety of meetings with the mission owner.

Based on the aforementioned task flow, it was determined that CPT planning activities center around four major tasks: identifying and requesting approvals to perform certain actions on networks/systems of interest, allocating trained operators to perform required tasks, obtaining access to various systems and networks identified as mission relevant cyber terrain, and communicating the mission plan via multiple meetings with various people (the team, mission owner, network points of contact, etc.). These tasks are critical to planning operations yet they are currently conducted with no specific planning tools. Instead, planners rely on spreadsheets, documents, pdfs, electronic calendars, training records, email, whiteboards, and sticky notes to organize and compile information. In consideration of the unique requirements within each mission, this process easily becomes tedious and timely for planners who often operate in time-constrained environments.

2.3 Iterative design phase

The information gathered during the analysis phase was transformed into a work aid prototype designed to facilitate the accomplishment of the four key CPT tasks. The planning process was streamlined through the integration of relevant information and the use of automation. In the iterative design phase of the human factors design process, multiple sessions were held with CPT planners and SMEs to receive feedback on the usefulness and usability of the work aid prototype.

3. Method of evaluation

In order to evaluate the effectiveness of the CPT planning work aid design, the prototype was evaluated against a baseline condition. The baseline condition consisted of the use of documents and spreadsheets to perform the four planning tasks described in the CPT Planner Workflow section (managing approvals, assigning resources, managing accesses, and scheduling meetings), mirroring the manner in which activities are currently performed. This condition is referred to as the manual condition due to its need for information to be manually integrated throughout task performance. One interest of this study was to observe the effects of the work aid on CPT members with no planning experience to perform the planning tasks efficiently and effectively. Therefore, this study had two independent variables: planning method (manual versus work aid) and planning experience (novice versus experienced). Fictitious missions were created for participants to plan during training and data collection. Dependent measures included time to complete each task, task performance score, and a subjective workload assessment using a Modified Cooper Harper Scale as shown in Figure 1. In addition to these three metrics, situation awareness (SA) probe questions (Endsley, 2000) were developed and asked after each task to determine if participants understood relevant information. An example of a SA question for the manage approvals task asked, “Do all of the actions necessary to complete this mission have prior approval?”. Time and accuracy of responses to all SA probe questions were recorded.

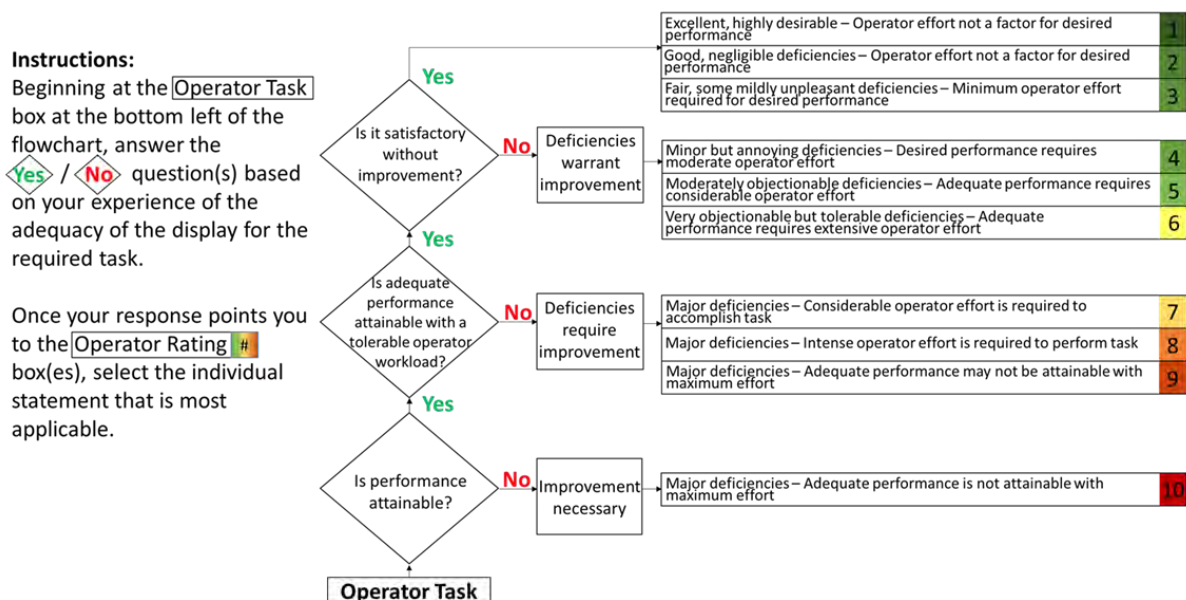


Figure 1: Modified Cooper-Harper Scale for Subjective Workload Assessment

3.1 Hypotheses

It was hypothesized that there would be a significant interaction between planning method and planning experience in that performance under the work aid condition would not be significantly different across various levels of planning experience, while performance in the manual condition would be significantly different with operator experience. In other words, participants with planning experience would perform better in the manual condition than their less experienced counterparts, while the work aid condition would facilitate statistically similar performance for participants of all experience levels. Additionally, it was hypothesized that there would be a positive correlation between task performance and SA question score, indicating that one should understand the relevant information if they performed well at the task.

3.2 Participants

Participants of the study included ten volunteers within the 837th Cyberspace Operations Squadron at Scott Air Force Base. These individuals were members of various CPTs at USTRANSCOM. Four participants reported having prior CPT planning experience while six did not. Participants completed the study on a voluntary basis with no other incentive than the ability to contribute to scientific research. The study was conducted in a secluded and secure area in order to limit distractions and interruptions.

3.3 Procedure

Participants were assigned to the condition order (manual first versus work aid first) in a counterbalanced fashion. Prior to the start of each condition, participants completed a training and practice portion to ensure a basic understanding of what was expected from each task. For consistency, training was accomplished via a pre-recorded video explaining the procedures for accomplishing the mission for each condition and questions were encouraged throughout this phase. For practice, participants performed a shortened version of each of the four tasks to ensure an understanding of what they were to accomplish. Following the completion of training and practice, participants completed all four tasks within the respective condition and this procedure was then repeated with the other condition. Throughout all phases of practice and data collection, participants rated their perceived workload after completing each task. Following the work load rating, SA questions relevant to the particular task were completed.

4. Results

Results showed that there was no significant interaction between the two independent variables of planning method and planning experience for all five dependent measure (task time, task score, workload, SA time, and SA score). Planning experience had no significant main effects for the dependent measures. However, independent sample t-tests showed significant differences for planning method, results shown in Table 1. Task time, workload, and SA time revealed significantly smaller values for the work aid condition when compared to the manual condition ($p < 0.0001$), while SA score was significantly higher for the manual condition ($p = 0.0077$). While the work aid condition yielded higher task performance scores when compared to the manual condition, these results proved insignificant ($p = 0.1511$).

Table 1: Independent sample t-test results for planning method

	Work aid (M, SD)	Manual (M, SD)	Test Statistic	P-value
Task time (Secs)	314.6, 103.8184	1020.3, 179.3743	-11.5033	< 0.0001
Task score	33.5, 0.9718	32.9, 1.3703	1.1927	0.1511
Workload	4.4, 0.6992	16.3, 3.566	-7.6388	< 0.0001
SA time (Secs)	508.6, 132.8953	567.6, 116.46	-1.7331	< 0.0001
SA score	28.5, 1.841	28.6, 0.81	1.5138	0.0077

Correlations were analyzed between all metrics for both conditions. For the manual condition, SA score and SA time had a moderate negative correlation ($r = -0.4974$), while task time and task score had low positive correlation ($r = 0.3929$). SA time and task time revealed a moderate positive correlation ($r = 0.5079$). For the work aid condition, SA score and SA time had a moderate negative correlation ($r = -0.5519$), while SA score and task score had a moderately positive correlation ($r = 0.631$). SA time and task time showed a moderate positive relationship ($r = 0.6571$).

5. Discussion and conclusion

The results of this study revealed that participant performance while utilizing the work aid was consistently better than when using the manual condition. Participants conducted tasks in statistically significant less time with the work aid condition compared to the manual condition, with the average time being more than 11 minutes faster. SA time was also significantly faster for the work aid condition, resulting in nearly a minute of saved time while upholding significantly higher performance scores. The results are indicative of the work aid's ability to facilitate high performance while also saving time, avoiding the typical speed/accuracy trade-off that is often observed in social science experiments. This study also revealed lower levels of perceived workload for the work aid condition compared to that of the manual. With an average workload rating of 1.1 (compared to the manual condition's 4.4), it is clear that the integration of information and automated assistance provided through the work aid saved the operators mental strain as they conducted the tasks.

The correlative analyses throughout this study also pose encouraging results. Literature is clear that situation analysis is an important component of effective task performance yet the development of SA metrics is often challenging in research. Under this consideration, the objective measurement of participant SA was a significant goal of this study. The moderate positive correlation between task score and SA question score suggests that participants with a high understanding of relevant information are performing tasks effectively, thus indicating that the SA questions focused on appropriate information relevant to task performance.

It was hypothesized that this study would reveal a significant interaction between planning experience and planning method. It was expected that participants with planning experience would perform better in the manual condition compared to their non-experienced counterparts due to familiarity of the conditions, while the work aid would result in higher performance for those with and without experience. These expectations were violated as planner experience did not significantly interact with planning method. A plausible explanation for this violation could be due to the use of fictional missions to supplement the study. CPT planners rely on experiential information (what members are best at which tasks, what actions are typically pre-approved, etc.) to conduct their tasks. However, the advantage of this experience could have been mitigated in the research setting which presented fictitious team members, tasking orders, and missions for purposes of this study.

The operational environment of CPTs is complex and dictates rapid action in order to continuously defend crucial military networks. The criticality of CPT operations has the potential to be damning if not performed effectively and within the given time constraints. At an organizational level, the prescribed work aid has the potential to momentarily improve the overall Cyber Mission Force through its facilitation of increased performance and decreased task time. Given the recent emphasis placed on a highly capable cyber force to maintain global security and stability, advancements for CPTs may have great implications on an international scale. At an individual level, the work aid also simplifies the stressful undertakings that must be performed throughout the planning process on a regular basis. Due to its alleviation of perceived mental workload, the work aid can greatly reduce the occupational stress placed on operators, leading to higher levels of job satisfaction and increased productivity.

In conclusion, planning is a crucial first step in mission execution for CPTs that necessitates tools capable of integrating information from a variety of sources. Automation is needed to minimize unnecessary work and promote timely, effective action that leads to mission success. The work aid proposed throughout this study proved effective at providing these requirements and yielded extremely significant results in its analysis. If implemented throughout Cyber Protection Teams, enhanced operator performance has the potential to impact cyber security on a global scale.

Acknowledgements

This material is based upon work supported by the Cyber Technologies Office of the Under Secretary of Defense Research and Engineering.

References

- Anderson, G. (2018) Cyber Protection Team Knowledge Acquisition Session. 11 July.
- Endsley M.R., (2000). "Direct measurement of situation awareness: validity and use of SAGAT" Situation Awareness Analysis and Measurement. Lawrence Erlbaum Associates Publishers, Mahwah, pp 147–173.
- Lathrop, S.D., Trent, S. and Hoffman, R. (2016) Applying Human Factors Research Towards Cyberspace Operations: A Practitioner's Perspective. *Advances in Intelligent Systems and Computing*, 501, pp 281–293.
- Pomerleau, M. (2019) Cyberwarriors lack planning tools. That could change. *Fifth Domain*, [online], <https://www.fifthdomain.com/dod/2019/11/25/cyber-warriors-lack-planning-tools-that-could-change/>.
- Schulz, A.E., Kotson, M.C., & Zipkin, J.R. (2015). *Cyber Network Mission Dependencies*. Lexington, Massachusetts: Lincoln Laboratory.
- Trent, L. S., Hoffman, R., Leota, T., Frost, C. R., & Gonzalez, M. D. (2016) "Cyberspace Operations and the People Who Perform Them." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60, No. 1, pp 216-217.
- United States. (2011) *Air Force Doctrine Publication 3-12 Cyberspace Operations*. Maxwell Air Force Base: U.S. Air Force.
- U.S. Army Cyber Command. (2020) DOD FACT SHEET: Cyber Mission Force, [online], <https://www.arcyber.army.mil/Info/Fact-Sheets/Fact-Sheet-View-Page/Article/2079594/dod-fact-sheet-cyber-mission-force/>.
- U.S. Congressional Research Service. (2020) *Defense Primer: Cyberspace Operations Overview*, [online], Congressional Research Digital Collection, <https://fas.org/sgp/crs/natsec/IF10537.pdf>.

I Know You by Heart: Biometric Authentication based on Electrocardiogram (ECG) signals

Christoph Lipps¹, Lea Bergkemper¹, Jan Herbst¹ and Hans Dieter Schotten^{1,2}

¹German Research Center for Artificial Intelligence, Intelligent Network Research Group, Kaiserslautern, Germany

²University of Kaiserslautern, Division of Wireless Communication and Radio Positioning, Kaiserslautern, Germany

Christoph.Lipps@dfki.de

Lea.Bergkemper@dfki.de

Jan.Herbst@dfki.de

Hans.Dieter.Schotten@dfki.de

Abstract: Trust, confidence and trustworthiness are of fundamental importance in human societies, usually established and sustained through personal relationships. But, due to the globalization and ongoing interconnection of everything up to Cyber-Physical Production Systems (CPPS) and the Internet of Everything (IoE), physical attendance is no longer necessary. (Remote) access to systems is possible from anywhere on the globe. Accompanied with the lack of personal relationship is the challenge to trust entities –humans or machines–, and proof the identity they are claiming to be. Whether it's payment transactions with smartwatches, logging in to systems, or accessing sensitive parts of buildings, the user's identity is the basic prerequisite. For human participants, for instance, the verification can be obtained through biometrics. These are distinguishable into physiological, biological and behavioral features, each characteristic but of varying difficulty to deduce them. Although using biometric features is not a new concept -indeed they are the oldest form of authentication-, modern approaches are shifting them back into focus. Improved sensor technology enables the identification of people by their gait, or to distinguish them by their characteristic gestures. This work highlights how the availability of (medical) data, and the possibilities of Artificial Intelligence (AI) contribute to the identification and authentication of humans. Therefore, Electrocardiogram (ECG) signals are recorded using a Microcontroller Unit (MCU) and ECG electrodes to derive a three-lead ECG. Using different Machine Learning (ML) algorithms: K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and Gaussian Naïve Bayes (GNB); it is analyzed whether the ECG signals are able to distinguish individuals. Thereby, the ML algorithms are compared with each other, determining which one achieves the best results. The results of the evaluation indicate that ECG signals are capable to distinguish humans based on their heartbeat in such a manner that they can be used as Human - Physically Unclonable Functions (Human-PUFs). Furthermore, the results give reason to assume that the algorithms can also be used for medical applications, for example to recognize heart diseases.

Keywords: Biometric Authentication; Human-PUFs; Physically Unclonable Functions; Physical Layer Security; Industrial Internet of Things; Trust

1. The Industrial Internet of Things and the matter of Authentication

The proliferation of broadband communications and the accompanying inter-connectivity: the coupling of everything with everything in Internet of Everything (IoE) and Industrial Internet of Things (IIoT) scenarios; is accelerated by recent developments in Artificial Intelligence (AI), the availability of large amounts of data and the ability to acquire, analyse and draw conclusions from this data. In combination with cloud computing, Cyber-Physical Production Systems (CPPS) and the virtualization of processes and systems, towards Digital Twins (DT) (Barricelli, et al., 2019), the analogue world is gradually being mapped into their virtual representative. In the process, however, the technological boundaries, places of operation and frontiers are blurring as well. There is no need of physical proximity to access a system, it can be accessed from anywhere in the world.

But this is also accompanied by strict requirements for the (cyber)security of these systems. The various scenarios –remote, on-site–, and entities involved –humans, machines, services–, though, pose major challenges. Although the fundamental requirements of any security concept are the same: the confidentiality integrity and availability (Schneier, 2015), the implementation is basically different. There is a difference between authenticating a human moving between several buildings and validating a machine statically located at one position in a factory or an Automated Guided Vehicle (AGV) moving independently in this factory. **Figure 1** shows this interactivity and the meshing of various components and demonstrates the difficulty of finding a holistic solution for these requirements.

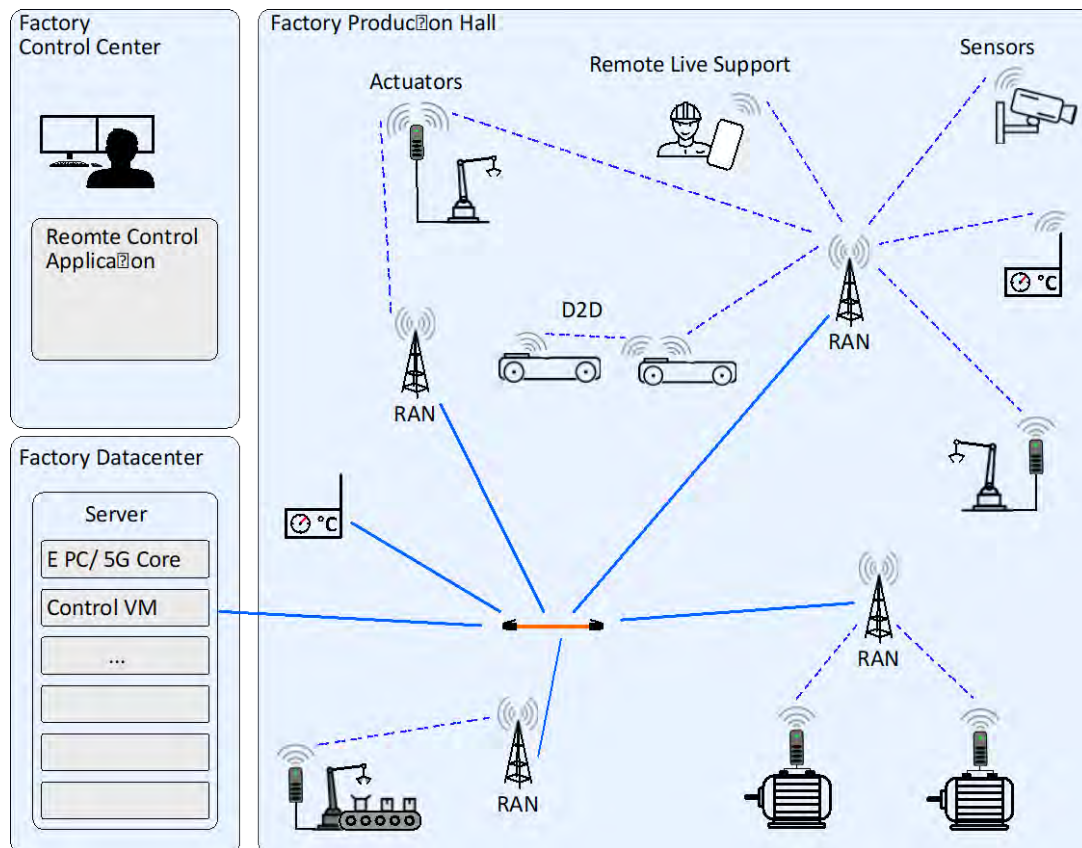


Figure 1: Industrial Internet of Things environment with wirelessly connected entities, including M2M, M2S and H2M applications

When it comes to Machine-to-Machine (M2M) and Machine-to-Service (M2S) applications, the approaches are focused on certificates, Hardware Security Modules (HSMs) and Trusted Platform Modules (TPMs), as well as encryption in general. Besides there are approaches with respect to Physical Layer Security (PhySec) aiming to utilize various physical characteristics of semiconductors (Lipps, et al., 2018) and the wireless propagation channel (Lipps, et al., 2020b) to authenticate devices and to derive cryptographic primitives. For all human-related scenarios, by contrast, biometric methods are (actually exclusively) appropriate. As the operating principle -the utilization of inherent/intrinsic characteristics- is analogous to PhySec, it is referred to as Human -Physically Unclonable Functions (Human-PUFs) (Lipps, Bergkemper & Schotten, 2021a).

The remainder of this work is organized as follows. In Section 2 an overall introduction to the authentication factors and their relevance in Multi-Factor Authentication (MFA) scenarios is given. Section 3 gives a brief insight into Artificial Intelligence and the global impact it enables, whereas Section 4 describes the setup for measuring the ECG values and the performance of the measurement sequences. A classification and discussion of the results is provided in Section 5. Section 6 concludes the work and provides an outlook on future work and the next steps to the further enhancements.

2. Biometric Authentication: An Evolution

The utilization of biometric features is not merely one of the simplest forms of authentication, it is actually the oldest form of identification. For instance, humans recognize each other by smell, voice, behaviour, stature, and facial features, even across distances and in groups, although only some or a few of the features are available. We recognize each other on the phone, although only phonetic and morphological features are present, which are additionally technically compressed and transmitted via (bad) channels. This uniqueness of biometric features is long known and have been used in forensics for more than 100 years, for instance in the context of fingerprints (Schneier, 2014).

But, increasing globalization and interconnectivity of environments and applications render this identification and authentication more challenging than ever before. Involved are no longer just humans in traditional Human-to-Human (H2H) scenarios, but also Human-to-Machine (H2M), Machine-to-Service (M2S), and all other forms

of mutual interaction. Furthermore, distances are also becoming wider, entities are no longer necessarily in close proximity and in the same area, remote access to systems is possible from anywhere on the globe. Altogether, the art of authentication changed, it grew more complex, while at the same time the technical possibilities increased significantly. In terms of biometric authentication, this implies that algorithms have significantly more difficulty in assigning human characteristics and detecting attempts of deception. However, modern approaches for biometric authentication are shifting them back into focus: methods of Artificial Intelligence (AI) enable the analysis of data and the ability to draw relevant conclusions from it; and improved sensor technology enables the identification of people by their gait, distinguish them by their characteristic gestures and by their heart rate.

In order to consider a technical solution of biometric authentication, it is reasonable to first understand the meaning of the term, which is why a short explanation is provided. Following the National Institute of Standards and Technology (NIST) authentication is “a process of establishing confidence” (Burr, et al., 2013). Combined with the proposal of O’Gorman as a “process of positively verifying the identity” of an entity (O’Gorman, 2003), an adequate indication can be made. “Establishing confidence” and “verifying the identity” form the basis of a trustworthy communication and exchange of information.

2.1 Authentication Factors: About Knowledge, Inherence and Possession

For the authentication of entities there are, as already mentioned, various methods, depending upon what should be authenticated in detail. On a first level, the methods are divided into three areas, as indicated on the left-hand side of **Figure 2**:

- Knowledge,
- Inherence, and
- Possession.

The *knowledge-based* factors encompass *things that someone knows*, such as passwords, Personal Identification Number (PIN) codes or secret patterns, whereas *inherence* factors describe *something that someone is*. These include factors based on physiological or biological characteristics. Factors *owned* by someone are referred to as *possession factors*. These are typically physical things such as key cards, token and metal keys.

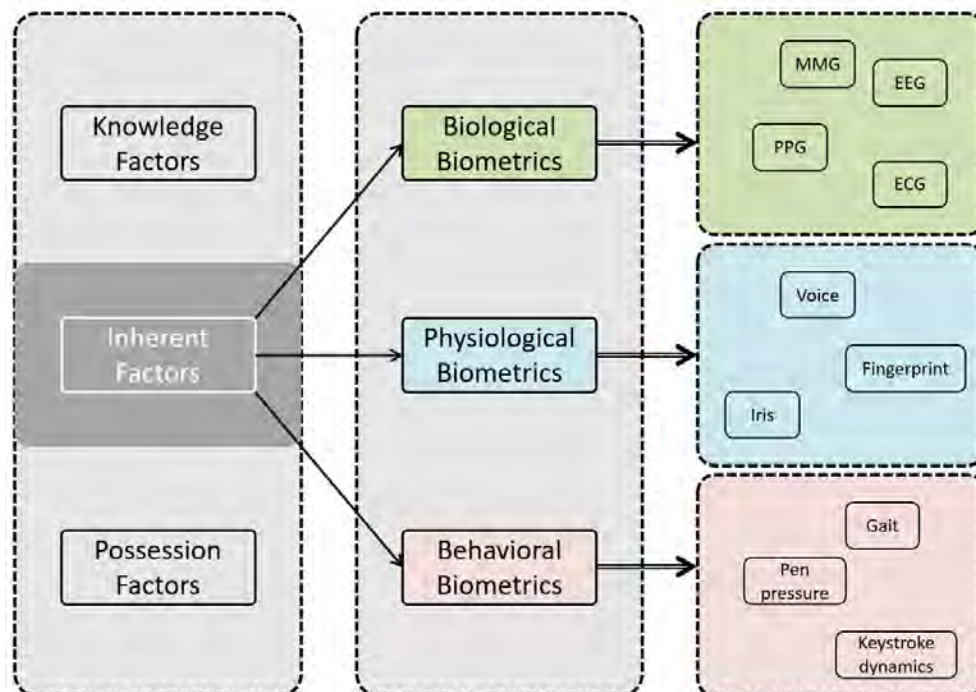


Figure 2: Authentication Factors are basically divided into three levels: Knowledge-, Inherence- and Possession Factors. Inherence Factors are further specified into the sub-groups: biological-, physiological- and behavioural biometrics, which could be further specified

Besides the mentioned categories there are approaches considering additional properties summarized by *context information*. This is additional information such as the geographical location, the Internet Protocol (IP) address and the status on an entity (Duque Anton, et al., 2018).

2.2 The Combination of Factors: Multi-Factor Authentication

As mentioned before, there are different groups of factors, which are differently suitable (or not) for individual applications. Moreover, there are applications with differently stringent security requirements. In its simplest form, this is just a single factor, such as a password or PIN code. This is called Single-Factor Authentication (SFA). With increasing demands on confidentiality and value of an access, the level of security can be increased by coupling such a password/PIN with a second factor to form a Two-Factor Authentication (2FA). Commonly, a factor from another group of factors is used for this purpose: if the first factor is *Knowledge*-based, for instance, the second factor should be either *Possession*-based or an *Inherence*-factors. As an example, this is the standard for withdrawing money at an Automated Teller Machine (ATM). The combination of bank/credit card (possession) and PIN code (knowledge) is required.

For a Multi-Factor Authentication (MFA), another additional factor is combined to obtain a *strong authentication*. Factors from each of the groups must be combined and validated together (Ometov, et al., 2018). To stick with the ATM example, withdrawing money could be made more secure by integrating an additional biometric factor into the authentication process. The possession of the card and the knowledge of the PIN is not a reliable validation of the identity of a person, it simply indicates that a person knows the PIN code associated with the card.

2.3 Human Characteristics: Biometrics

As discussed in Section 2.2, current biometric methods offer additional benefits in terms of further enhancing the level of security. Nevertheless, when using them it is important to keep in mind that they are “powerful and useful, but they are not keys” (Schneier, 1999). Unlike mechanical keys, the human body is a living organism, which is constantly changing: cells are dying and new ones are being formed (Renaud, 2005). The resulting deviations and inaccuracies of the parameters must be considered in a technical validation.

As indicated in **Figure 2**, biometrics are attributed to the “inherence” factors. As illustrated in the central block of the figure, these can in turn be differentiated into physiological, behavioural and biological characteristics (Yang & Nanni, 2011).

Also in **Figure 2**, the factors can be further subdivided into *biological biometrics* –representing the *inner part* of living beings –, *physiological biometrics* –morphological human factors–, and *behavioural biometrics*, referring to pattern of human behaviour (Jain, Ross and Pankanti, 2004). Thereby biological biometrics are, besides others, muscle contractions by deriving a Mechanomyogram (MMG) (Enamamu et al., 2017), eye movement by means of Electrooculography (EOG), volumetric changes of blood in peripheral using Photoplethysmography (PPG) (Vhaduri & Poellabauer, 2017), bio-electrical signals such as cerebral waves of an Electroencephalogram (EEG), or electric muscles activities of the heart via an Electrocardiogram (ECG) (Singh & Singh, 2012).

Physiological biometrics are morphological factors (Idrus et al., 2013) such as the characteristics of the voice, the pattern of the iris, the geometry of the hand and fingerprints. In contrast, gestures, gait, keystroke dynamics, lip motion, pen pressure, signature geometry and speed are attributes related to the (learned) behaviour - congenital, trained and culturally acquired-, and are ascribed to *behavioural biometrics* (Bolle et al., 2004).

As the classification already indicates, there are many different biometrics that are also suitable for authentication, this is also shown by numerous current research studies on this topic: *Neal and Woodard* provide a survey for mobile device security in which they compare several physical biometrics such as fingerprint, iris, face and periocular as well as factors such as gait, keystroke dynamics, voice, and touch gestures (Neal & Woodard, 2016). Besides, *Mahfouz et al.* describe behavioural biometric authentication based on smartphones by measuring the gait with smartphone integrated sensors (Mahfouz, Mahmoud and Eldin, 2017). A similar approach is taken by *Lipps, Herbst and Schotten* as they build a system to analyse the gait with a sensor matrix located in the insole of a shoe (Lipps, Herbst and Schotten, 2021b). Furthermore, *Rui and Yan* are highlighting potential risks, common network attacks and faking sensors by resubmitting biometric features. Their overview about existing biometric authentication systems include a rating system with respect to accuracy, efficiency, security and privacy features (Rui & Yan, 2018). With respect to MFA scenarios, *Hammad et al.* propose the

combination of ECG and fingerprint data, to obtain a multimodal biometric system (Hammad, Liu and Wang, 2018).

However, this line-up also indicates, that “biometrics are unique identifiers, but they are not secrets” (Schneier, 1999). People leave their biometric features such as fingerprints in many places, and some, such as the iris, facial characteristics or body shape, are easy to recognize even from the outside. Moreover, the loss of a factor, whether by theft (digital representation, validation feature) or by destruction (loss of a finger, change of friction grooves), implies the destruction and unusability of the feature forever.

But nevertheless, cryptographically, these biometric features are generally stronger than standard passwords; for example, a fingerprint contains significantly more information than a 128-bit passcode and given their always with you character -biometric features are always directly linked to the person to be authenticated- it is almost impossible to lose or forget such a biometric feature. The challenge in using it is twofold: i) validating that a factor belongs to the right person, and ii) that the factor matches the stored factor (Schneier, 1999), but this is where modern approaches such as Artificial Intelligence can assist.

3. Artificial Intelligence and its Applications

Alan Turing provided the so called Turing Test back in 1950 to provide a “[...] satisfactory operational definition of intelligence” (Russell & Norvig, 2016). Till this time there is no known complete AI which is able to satisfy all aspects of this test. In today’s world in nearly every aspect of the modern life AI-techniques have at least some kind of approach. Be it in the fields of online shopping, advertising, enhanced Web searches, or in the topics of transportation, efficient manufacturing, or in the fields of food and farming. There is a huge variety of AI-subfields, ranging from general learning and recognising algorithms to specific ones for playing chess, proving mathematical theorems, writing poetry, driving cars, or for diagnostics of different diseases (Russell & Norvig, 2016). AI techniques encompass intelligent algorithms capable of performing certain tasks, sometimes better than a normal human would.

AI is used to create useful systems which are able to decide “intelligent” on the basis of the given data (Alpaydin, 2016). Therefore we use the continuous mathematics developed and discovered through the last centuries. Already in 1936 Turing showed that every possible computation can in principle be performed with a mathematical system. It was called the universal Turing machine (Russell & Norvig, 2016). At 1985 the thinking went one step further and it was said that “thinking and computing are radically the same thing” (Haugeland, 1985). Today state-of-the-art image or handwrite-recognise-systems are nearly as good at recognising specific objects or numbers as a normal human could do, in a much faster amount of time (Dodge & Karam, 2017).

According to *Margaret Boden* (Boden, 2016) there are five major types of AI, namely: Classical/Symbolic; Artificial Neural Networks; Evolutionary Programming; Cellular Automata and Dynamical Systems. The *Classical/Symbolic AI*, especially when combined with statistics, has its approaches in model learning, pattern detection, planning and reasoning. *Artificial Neural Networks* are especially good at pattern-recognition. The power of this system is that it finds necessary features by its own, although data is absolutely required. This data used will develop a system on which a trained algorithm will base its decisions. How wrong this can go showed the known amazon AI recruiting tool (Lauret, 2019) which preferred the male gender for recruiting, first without a clear reason. *Evolutionary Programming* is used to throw light on the biological evolution and brain development whereas *Cellular Automata* describes the field of discrete mathematical systems to characterize local interactions and dynamical computational evolutions. *Dynamical Systems* on the other hand can be used to model the development in living organisms.

For the approach of this paper the classical/ Symbolic AI was used to detect certain patterns. The Classical ML has three broad types: supervised; unsupervised and reinforcement learning (Boden, 2016). For supervised learning the system knows the patterns it has to predict. The system can be trained directly on the data. A system is then able to recognise a specific pattern, with the right selection of features at the training stage. Different mathematical algorithms and theories explored in the last centuries can be used for this approach. Unsupervised learning on the other hand means the system does not have a certain output, but tries to cluster and to find patterns with the given Data and features. Reinforcement learning means that a system gets trained with the help of rewards on certain training stages.

In approach of this work the system is trained with the help of supervised learning. Therefore the classification algorithms: K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and Gaussian Naïve Bayes (GNB) are used and compared. The KNN classifier algorithm determines the distance between different points in a diagram. The closest datapoint to other points gets then classified as an associated class. The Letter “K” indicates how many neighbors are used for determining the distance. The SVM classifier algorithm uses supporting boundaries -vectors- between the point of certain classes -the hyperplanes-. The algorithm then tries to maximize the margin between the hyperplane and the supporting vectors which are at the position of certain data points. The GNB classifier algorithm is based on the Bayesian statistics and therefore with condition probabilities for not independent events. The GNB is the extension estimating the mean and standard deviation for certain training data (Berrar, 2019).

4. Measuring ECG Signals: The Experimental Setup

In terms of biometrical authentication, the ECG can be classified as a biological factor, since the signal is a visualization of the electrical activity of the heart. An advantage over physiological features is its robustness against any physiological changes. Simple everyday accidents and procedures such as cuts on the fingers and swelling of the face due to medical treatments would normally be problematic for fingerprint scanning methods or face recognition. With the rapid evolution of modern technology, accessibility of heart rate supervision is made possible with wearables such as smartwatches. Due to the characteristics of the ECG signal, a continuous authentication, desirable for the priority mentioned environment is possible.

The signal itself is a recording of the electrical activity of the heart, measured on the skin. It is observable with every heartbeat, consisting of the mechanical contraction of one heart chamber and the following relaxation. The resulting signal of one heartbeat is called a PQRST Wave, as depicted in **Figure 4**. The measurement can be done using either one or up to twelve electrodes, which need to be placed on specific points of the body. To reduce the complexity of the setup and improve the applicability in real life, a three-lead ECG is used for the experiment, which is a common setup for long-term measurements. Between each of the electrodes, the voltage is measured resulting in a vector according to the bipolar Einthoven-derivations.

For the recording and processing of data an Arduino Mega ADK Micro Control Unit (MCU) is connected to the ECG Click Sensor from MikroElektronika. The measured data is subsequently transmitted to the Arduino, which forwards it to the computer, where further processing and calculations are conducted using the programming language Python. Regarding a proof-of-concept data of 25 subjects are recorded, out of which 12 are male and 13 are female. In order to increase the integrity of the results similar conditions for the measurements of the individuals had to be ensured which is why the following aspects were considered. A challenging aspect regarding the recording process is the sensitivity of the cables and sensors. To ensure usable recordings, the data can only be acquired without the person moving. Furthermore, other environmental factors often cause noise. Those are for the instance the charging of the computer during the data acquirement. To avoid this, the subjects were measured twice for a duration of 5 minutes without prior exercise to achieve a resting heart rate during the measurement. Additionally, movement is kept to a minimum. For hygiene and reliability purposes, new pads were used for each individual. Further processing is done on a desktop PC following the steps shown in **Figure 3**.

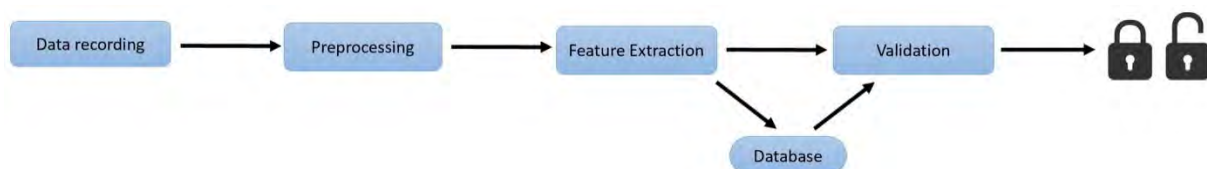


Figure 3: Steps of processing of the ECG data. After the data has been recorded, it is subjected to appropriate pre-processing and feature extraction.

5. Classification and discussion of the results

To validate the identity of an individual based on the ECG recordings, a classification of the data is necessary. In order to apply ML approaches, the recordings are pre-processed. In a first step, a filter is applied to reduce noise caused by the environment or movement of the person. Subsequently the recordings are segmented by finding characteristic waves, the so called PQRST Waves. One of these waves including its characteristic features are

depicted in **Figure 4**. Each segment includes 6 waves in order to counteract outliers and anomalies in the filtered data.

To keep the required storage space to a minimum and enable a comparison, features are extracted from each segment and stored in a feature vector. The applied fiducial feature extraction focusses on significant points in one PQRST wave. The resulting feature vector for one segment includes the medium values of the PQRST wave duration, R peak amplitude, T peak amplitude, QRS wave duration, P peak amplitude, Q peak amplitude, S peak amplitude and medium value of all data. These vectors are used as an input for the classification algorithms.

Comparing the measured data from 2 different subjects, significant differences are observable. **Table 1** includes two vectors of the same person A, extracted from distinct measurements and one feature vector from a different subject B. It is observable, that the duration of the QRS Wave does not differ significantly between the person A and B, while the amplitudes of the S Peaks are recognizably different. However, divergences also occur within the results of one subject. This becomes obvious looking at the different durations of the PQRST Wave in the given feature vectors from subject A. This observation is confirmed by **Figure 5** and **6**, with the left one showing two recordings of subject A, while the right one depicts one ECG signal of Subject A and one of B.

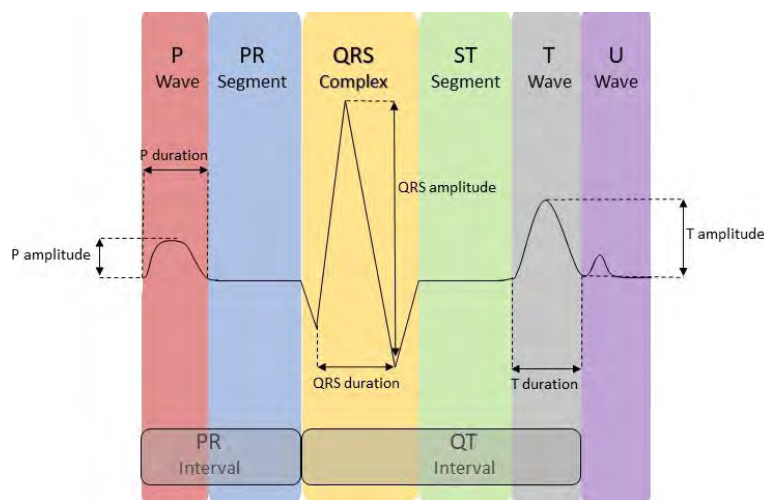


Figure 4: Characteristics of a PQRST Waves

Table 1: Feature Vectors of Recordings of different Subjects

Subject	PQRST Wave Duration	R Peak Duration	T Wave Amplitude	QRS Wave Duration	P Peak Amplitude	Q Peak Amplitude	S Peak Amplitude	Medium Value of all Data
A	661.44	639.66	286.18	102.92	220.14	163.04	79.80	196.86
A	625.70	641.34	279.39	102.92	217.46	162.54	77.79	196.31
B	730.67	448.38	236.45	102.19	214.06	190.04	153.55	205.28

Another challenge for the comparison are significant differences in the structure of the PQRST Wave of each individual person. They can differ in the number of occurring waves, as the S or P Wave is not always included, the orientation and order of the amplitude compared to the other waves.

Due to this drawbacks, three different ML approaches, namely the KNN Classifier, the GNB Classifier and the SVM, are applied, as they provide a higher level of adjustment to the data. Table 2 shows a comparison of the accuracies of these classification methods, depending on the database size.

Table 2: Comparison of the Classifiers

Number of Subjects	KNN Classifier	SVM	GNB Classifier
1	100%	100%	100%
2	100%	100%	100%
5	76.7%	97%	77%
10	62.3%	94%	61%
15	72%	94%	77%
20	37%	92%	90%

Number of Subjects	KNN Classifier	SVM	GNB Classifier
25	47%	94%	89%

The first approach is the KNN algorithm, which assigns the vectors to certain classes based on their distance to the clusters' centroids. This approach works for smaller sets, as for instance for two different persons a clear separation between the classes can be observed. However, problems occur on a larger database, which is caused by the higher rate of overlapping classes. This higher number of misclassified vectors can among others be reduced to the underlying distance measure, which is not adjusted to the specific properties of the data.

In comparison, the SVM is applied for the given set of data. The results of the SVM are more accurate than the ones for the k means algorithm, since the hyperplane, separating the different classes from each other, is more adjustable to the data. In the opposite, the distance measurement used for the KNN algorithm cannot respect outliers in the same degree. However, outliers and anomalies are also a challenge for the SVM, as the hyperplane might not be able to adjust all vectors to the right classes. However, the accuracies for the different database sizes show, that only few vectors are misclassified. In addition, the wrongly classified sample are recordings from different persons, where only one out of them is wrongly classified. Therefore, the possibility of wrongly accepting or rejecting a user in a real application is rather small. Although the percentage of the correct classified persons get less with a greater database, it is for a size up to 25 persons constantly over 90%. The third tested classification method is the GNB Classifier. Compared to the results for 25 individuals, it can be observed, that the number of misclassified points decreases with more people. Thus, the overall accuracy increases. Furthermore, only isolated outliers can be observed for 25 subjects, whereas results from 10 people show more misclassified data for the same person. In addition, the results for 10 persons show one subjects, where nearly all samples are classified wrongly. This is an unsatisfying result, since it would cause a DoS for person in an established authentication system or would grant access to the wrong person.

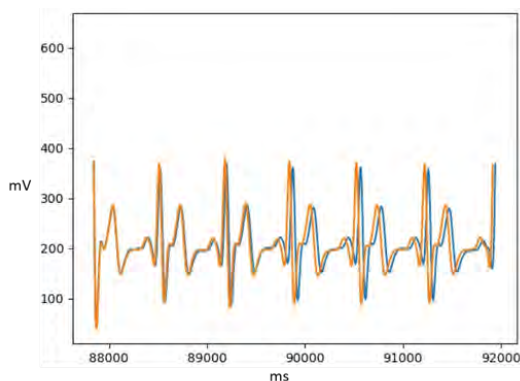


Figure 5: ECG Signals of one Individual

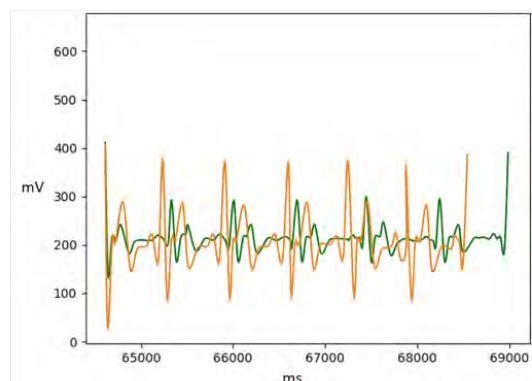


Figure 6: ECG Signals of two Individuals

Considering the feature vectors of these two subjects, it is revealed that all entries of the vectors only slightly differ from the other person's ECG signal. This shows, that although the GNB considers each feature individually, it gets problems with small datasets and similar samples of two subjects. However, for a larger set of data, smaller differences in the vectors gain importance. Hence, the GNB classifier provides a more accurate result for the data set including data from 25 people.

Compared to the KNN algorithm, the GNB Classifier achieves rather similar accuracies for smaller sets of data, while for a larger database, it outperforms the clustering algorithm significantly. The SVM achieves the best and most stable results for all database sizes given in the table. However, for a set of 25 people, the GNB classifier nearly achieves the same accuracy. This observation can be reduced to the underlying algorithms. As mentioned before, the Euclidean Distance is used in the KNN algorithm as a measurement for the differences, whereas the measurement of the other two approaches are adjusted to the data. Thus, its performance is less satisfying. In addition, a disadvantage of the KNN algorithm are the problems with the classification of non-globular shaped data clusters. Compared to this, the GNB classifier considers the Gaussian Distribution for each individual feature, comparing all values of the vector separately from each other. However, it assumes, that all features are independent of each other, which cannot be satisfied, as the duration of the overall wave and the smaller waves it includes are obviously depending. The SVM puts the focus on the interaction of the values instead of comparing them separately from each other. Therefore, it outperforms the GNB Classifier when comparing two

samples with similar values at specific entries in the feature vector, as a combination of all values is used for the comparison. Thus, the SVM results in the highest accuracy, considering the given database including the samples of 25 persons and the chosen separation of trainings and test data. However, these conclusions are made considering a rather small data set. To determine the accuracies of a larger database, tests have yet to be conducted.

6. Outlook and Future Work

The use of biometric factors for the authentication of human entities is shifting back into focus due to current possibilities and modern sensor technology. Bygones are the days of simple fingerprints or iris scanners, especially since many of these approaches are no longer secure. Methods such as gait recognition, behaviour-based identification and the utilization of Electrocardiogram values are of growing relevance.

For this purpose, this work describes a setup and proof-of-concept for ECG-based authentication. With a simple test-set-up based on Arduino Microcontroller, ECG click shield and ECG-pads the values of 25 persons are recorded, processed and evaluated. Three different Machine Learning methods -K-Nearest Neighbor, Support Vector Machines and Gaussian Naïve Bayes-, were used and compared with respect to their suitability and performance. The results show that for only a few individuals all algorithms perform equally well, which was to be expected. As the number of participants increases, only the SVM delivers very good results and the GNB classifier still delivers good results. After a dent in performance, however, the KNNs are also improving again. The work thereby demonstrates that ECG-based authentication works in principle and can be used as a factor for authentication.

Due to the simple test setup, only ECG values with persons at rest could be recorded. In the next step, with improved hardware, persons will also be examined during other activities such as walking or light and medium physical exertion. It is expected that the heart values are even more significant. Improved hardware also includes the use of a 1-lead ECG, instead of the 3-lead ECG used in this case, for example with a sensor on the wrist.

Because the ML methods used performed differently, it is necessary to investigate whether a combination of the methods might provide better results. In addition, other ML methodologies will also be included from the areas supervised-, unsupervised, and reinforcement learning.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KISK003K Open 6G Hub). The authors alone are responsible for the content of the paper.

References

- Alpaydin, E., "Machine Learning: The New AI", *The MIT Press Essential Knowledge Series*, ISBN: 978-0262529518, 2016.
- Barricelli, B.R., Casiraghi, E., and Fogli, D., "A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications", *IEEE Access*, Issue 7, pp. 167653 -- 167671, DOI: 10.1109/ACCWSS.2019.2953499, 2019.
- Berrar, D.P., "Bayes' Theorem and Naive Bayes Classifier", *Encyclopedia of Bioinformatics and Computational Biology*, DOI: 10.1016/B978-0-12-809633-8-20473-1, 2019.
- Boden, M.A., "AI - Its nature and future", *Oxford University Press*, ISBN: 978-0198777981, 2016.
- Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., and Senior, A.W., "Guide to Biometrics", Springer Professional Computing, Springer, DOI: 10.1007/978-1-4757-4003-3, 2004.
- Burr, W.E., Dodson, D.F., Nweton, E.M., Perlner, R.A., Polk, W.T., Gupta S., and Nabbus, E.A., "Electronic Authentication Guideline - NIST Special Publication 800-63-2", National Institute of Standards and Technology, DOI: 10.6028/NIST.SP.800-63-2, 2017.
- Dodge, S., and Karam, L., "A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions", *Cornell University - Computer Science - Computer Vision and Pattern Recognition*, arXiv:1705.02498, 2017.
- Duque Anton, S., Fraunholz, D., Lipps, C., Alam, K., and Schotten, H.D., "Putting Things in Context: Securing Industrial Authentication with Context Information", *International Journal on Cyber Situational Awareness (IJCSA)*, vol. 3, no. 1, DOI:10.226119/IJCSA.2018.100122, 2018.
- Enamamu, T.S., Clarke, N., Haskell-Dowland, P., and Li, F., "Transparent authentication: Utilising heart rate for user authentication", *12th International Conference for Internet Technology and Secured Transactions (ICTST)*, Cambridge, United Kingdom, DOI: 10.23919/ICTST.2017.8356401, 2017

- Hammad, M., Liu, Y., and Wang, K., "Multimodal Biometric Authentication Systems Using Convolution Neural Network Based on Different Level Fusion of ECG and Fingerprint", *IEEE Access*, Issue 7, pp. 26527-26542, DOI: 10.1109/ACCESS.2018.2886573, 2018
- Haugeland, J., "Artificial Intelligence: The Very Idea", *The Massachusetts Institute of Technology*, ISBN: 978-0262081535, 1985.
- Idrus, S.Z.S., Cherrier, E., Rosenberger, C., and Schwartzmann, J.-J., "A Review on Authentication Methods", *Australian Journal of Basic and Applied Sciences*, 7(5), pp. 95-107.
- Jain, A., Ross, A., and Pankanti, S., "Biometrics: a tool for information security", *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125-143., DOI: 10.1109/TIFS.2006.873653
- Jain, A., Ross, A., and Prabhakar, S., "An Introduction to Biometric Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20, DOI:10.1109/TCSVT.2003.818349, 2004
- Lauret, J., "Amazon's sexist AI recruiting tool: how did it go so wrong?", [Online]
Available at: <https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e>, 2019, [Zugriff am 30 10 2021].
- Lipps, C., Ahr, P., and Schotten, H. D., "The PhySec Thing: About Trust and Security in Industrial IoT Systems", *Journal of Information Warfare*, vol. 19, no. 3, pp. 35-49, 2020a
- Lipps, C., Bergkemper, L., and Schotten, H. D., "Distinguishing Hearts: How Machine Learning identifies People based on their Heartbeat", *6th International Conference on Advances in Biomedical Engineering (ICABME2021)*, DOI:10.1109/ICABME53305.2021.9604855, 2021a
- Lipps, C., Herbst, J., and Schotten, H. D., "How to Dance Your Passwords: A Biometric MFA-Scheme for Identification and Authentication of Individuals in IIoT Environments", *16th International Conference on Cyber Warfare and Security (ICWS)*, Cookeville, TN, USA, 2021b.
- Lipps, C., Mallikarjun, S.B., Strufe, M., Heinz, C., Grimm, C., and Schotten, H.D., "Keep Private Networks Private: Secure Channel-PUFs, and Physical Layer Security by Linear Regression Enhanced Channel Profiles", *3rd International Conference on Data Intelligence and Security (ICDIS)*, South Padre Island, TX, USA, DOI: 10.1109/ICDIS50059.2020.00019, 2020b
- Lipps, C., Weinand, A., Krummacker, D., Fischer, C., and Schotten, H.D., "Proof of Concept for IoT Device Authentication Based on SRAM PUFs Using ATMEGA 2560-MCU", *1st International Conference on Data Intelligence and Security*, South Padre Island, TX, USA, 2018, DOI: 10.1109/ICDIS.2018.00013
- Mahfouz, A., Mahmoud, T.M., and Eldin, A.S., "A survey on behavioral biometric authentication on smartphones", *Journal of Information Security and Applications*, vol. 37, pp. 28--37, 2017, DOI:10.1016/j.jisa.2017.10.002.
- Neal, T.J., and Woodard, D.L., "Surveying Biometric Authentication for Mobile Device Security", *Journal of Pattern Recognition Research*, vol. 11, no. 1, pp. 74--110, 2016, DOI: 10.13176/11.764.
- O'Gorman, L., "Comparing passwords, tokens, and biometrics for user authentication", *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021--2040, DOI: 10.1109/JPROC.2003.819611.
- Ometov, A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., and Koucheryavy, Y., "Multi-Factor Authentication: A Survey", *cryptography*, MDPI, vol. 1, no. 2, 2017, DOI: 10.3390/cryptography2010001.
- Renaud, K., "Evaluating Authentication Mechanisms" *Security and Usability: Designing Secure Systems That People Can Use*, O'Reilly, pp. 103-128.
- Rui, Z., and Yan, Z., "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification", *IEEE Access*, Issue 7, pp. 5994-6009, DOI:10.1109/ACCESS.2018.2889996.
- Russell, S.J., and Norvig, P., "Artificial Intelligence: A Modern Approach", *Pearson*, 3rd Edition, Addison Wesley, 2009, ISBN: 978-0136042597
- Schäfer, G., and Rossberg, M., "Security in Fixed and Wireless Networks", 2nd Edition, Heidelberg, Germany, John Wiley & Sons Ltd., 2016, ISBN: 978-1119040743.
- Schneier, B., "The Uses and Abuses of Biometrics" *Communications of the ACM*, vol. 42, no. 8, p. 136.
- Schneier, B., "Carry on: Sound Advice from Schneier on Security", 1st Edition, Indianapolis, Indiana, United States, John Wiley & Sons Inc., 2014, ISBN: 978-1118790816.
- Schneier, B., "Applied Cryptography: Protocols, Algorithms and Source Code in C", 20th Anniversary Edition, Indianapolis, IN, USA, John Wiley & Sons Ltd., 2015, ISBN: 978-1-119-09672-6.
- Singh, Y.N., and Singh, S., "Evaluation of Electrocardiogram for Biometric Authentication", *Journal of Information Security*, vol. 3, no. 1, 2012, DOI:10.4236/jis.2012.31005.
- Vhaduri, S., and Poellabauer, C., "Wearable device user authentication using physiological and behavioral metrics", *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Montreal, QC, Canada, 2017, 10.1109/PIMRC.2017.8292272.
- Yang, J., and Nanni, L., "State of the art in Biometrics", 1st Edition, London, United Kingdom, IntechOpen Limited, 2011, ISBN: 9789535155669.
- Zhou, B., Singh, M.S., Doda, S., Yildirim, M., Chengm J., and Lukowicz, P., "The Carpet Knows: Identifying People in a Smart Environment from a Single Step", *IEEE International Conference on Pervasive Computing and Communications (PerCom-17) - First International Workshop on Pervasive Smart Living Spaces*, Kona, Hawaii, United States, 2017, 10.1109/PERCOMW.2017.7917618.

Shallow Deep Learning using Space-filling Curves for Malware Classification

David Long and Stephen O'Shaughnessy

Technical University Dublin, Ireland

davielong86@gmail.com

Stephen.OShaughnessy@tudublin.ie

Abstract: The incidents of malware attacks are continually increasing at a rapid rate, thanks to the lucrative potential in schemes such as ransomware, credential stealing Trojans and cryptominers. Their explosive growth is compounded by the ease with which variants can be created from original strains. As a result, anti-virus organisations are struggling to keep up, with some reporting upwards of 14 million samples processed per month. These sheer volumes have caused a shift towards machine learning and artificial intelligence in an effort to alleviate the manual burden of analysis and classification. This research presents a novel framework for the classification of malware into distinct family classes through computer vision and deep learning. In the proposed framework, malware binaries are represented in an abstract form as images mapped through mathematical constructs known as space-filling curves. Convolutional neural networks were constructed and applied to the malware images to build predictive models for classification. The models were optimised using an auto-tuning function for the hyper parameters, which included Bayesian Optimisation, Random search and HyperBand, providing an exhaustive search on the hyper parameters. On a training dataset of 13k malware samples from 23 distinct families, the models yielded an average score of 95% for precision, recall and f1-score. The final deep learning model was validated for robustness against a dataset of more recent variants, comprising 12,816 samples from 16 malware families, returning classification scores of 95%, 86% and 90% for precision, recall and f1-score. The final model was demonstrated to outperform a similar benchmark model considerably. The results show the potential of the deep learning framework as a viable solution to the classification of malware, without the need for manually intensive feature generation or invasive processing techniques.

Keywords: Malware image classification, Deep Learning, Computer Vision, Space-filling curves, H-curve

1. Introduction

The continuing global upsurge in malware attacks can be greatly attributed to the potential profits to be gained from schemes such as ransomware and credential stealing Trojans. Anti-virus companies, such as AV-Test, report over 450,000 new malicious programs processed daily (AV-Test, 2021). Analysis on this scale is not manually feasible and so analysts and researchers constantly strive to produce more scalable solutions. More and more, researchers are utilising machine learning as a means of alleviating the manual analysis burden of malware classification. Typically, features are collected and then are passed to classification algorithms to train models for prediction, so feature generation, i.e., the extraction and subsequent choice of features, is crucial to the success of the model. Broadly speaking, features can be extracted in two ways: statically (from non-running binaries) or dynamically (from binaries in execution). Common static features are in the form of frequency or sequence data, such as n-gram or PE file header analyses (Khalilian et al., 2018; Kolter and Maloof, 2004). Dynamic features are extracted by executing malware and recording the behaviour by monitoring its interactions with the operating system, for example through API call sequences and graph representations (Fukushima et al., 2010). However, these approaches are not without their limitations. Frequency or sequence-based methods are context dependent, so any word or frequency contexts outside of a given range are not recognised; dynamic methods are prone to producing large volumes of false positives as much of the behaviours recorded are common to benign applications also.

Computer vision and image processing techniques have been presented as an alternative to the more typical approaches to malware classification (Wagner et al., 2015). Under these approaches, binary code is mapped or transformed into image formats, such that the images retain characteristics that can be extracted as discriminant features for identification and classification. The benefit of image representations is that the conversion process is less invasive than other methods of analysis such as those discussed previously, which simplifies feature generation. This paper presents a novel method of malware family classification that combines computer vision and deep learning through convolutional neural networks (CNNs). Malware binaries are mapped to a 2-dimensional image format through space-filling curves (SFCs). SFCs are mathematical constructs whose ranges contain the entire 2-dimensional unit square i.e., the 2-dimensional unit square represents an image of $n \times n$ pixels. SFCs traverse through every pixel point of a regular spatial region, such that the spatial locality of the data

is preserved. This is significant, as it enables the internal structure of the original malware binaries to be retained and represented in the resulting images, which can then be used for classification purposes.

In summary, this research makes the following contributions:

- *A novel approach to malware image classification through the application of Deep Learning & Computer Vision to malware Space-filling curve images.*
- *Comparative analysis with benchmark previous work to determine viability of proposed method.*
- *A pre-processed labelled malware dataset consisting of approximately 26k samples in space-filling curve image format.*

2. Related work

The visualisation of binary files as images was introduced in the work of Conti et al. (2008), whereby a method of reverse engineering binary files into visual images, dubbed "byteplots", was presented to enhance the capabilities of text-based hex editors. In this work, Conti et al. derived a visual taxonomy of binary file fragments using byteplots, to aid in the identification of common file formats. Conti et al. (2010) extended their previous work by applying machine learning to classify file types by their visual features. On a dataset of 14k samples comprising 14 file types, including machine code, encrypted data and plain text, the authors reported high classification scores, particularly from unencoded files, where accuracies of 98.7% and above were recorded. While the initial work presented by Conti et al. did not focus on malware classification, it demonstrated that the internal static structure of binary files could be represented as 2-dimensional images, which fostered further work in the application of byteplots to malware classification.

Previous works in image-based malware classification have focused predominantly on the byteplot representation for image-based malware classification. Nataraj et al. (2011) were the first to apply the byteplot mapping in this context. The authors derived feature vectors using GIST filters tuned to varying scales and orientations. They evaluated their method using a dataset, dubbed Malimg, comprising 9,548 samples from 25 malware families, reporting an average classification accuracy of 92%. Vasan et al. (2020) presented a method of transfer learning using the established pretrained VGG16 and ResNet-50 architectures with multiclass SVMs as classifiers, applied to the Malimg dataset for malware family classification. The authors reported an average accuracy score of 99.5% on their ensemble method, which out-performed 14 other similar methods. Le et al. (2018) used 6 multi-layered convolutional neural networks on binary files from the Kaggle malware dataset (Kaggle, 2015) that were converted to grayscale images. Results on the data showed 98.8% classification accuracy.

In contrast, limited works exist on SFCs in the image-based malware classification domain. This paper extends the work of O'Shaughnessy (2019), where the efficacy of space-filling curves was evaluated as a means of representing malware variants for classification. Three SFC image datasets were produced by mapping 9,235 32-bit executable malware samples from 28 distinct families to images using Z-order, Hilbert and Gray-code curve traversals. Features extracted via Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and Gabor filters were used to train Random Forest, K-Nearest Neighbour and Support Vector Machine classifiers. The best results were obtained from the KNN-HOG model trained on the Z-order dataset, with precision, recall and accuracy metric scores of 94.5%, 87.1% and 91.6% respectively. A comparative assessment with the method presented by Nataraj et al. (2011), showed the KNN-HOG Z-order SFC model outperformed the GIST byteplot method against previously unseen samples.

This research addresses several of the limitations associated with previous works. First, the framework utilises SFC images to classify malware. SFCs have previously demonstrated better classification performance compared to the predominant byteplot method (O'Shaughnessy, 2019). Second, the malware datasets used in this research were compiled from recent repositories which are representative of current malware variants, compared with the Malimg and Kaggle datasets, which were compiled in 2011 and 2015 respectively. Finally, this research explores the use of deep learning applied to SFC images for malware classification, which has not been documented previously.

3. Methodology

The research presented in this paper provides a study of deep learning and space-filling curves as a means of classifying malware into their respective family classes. The method of the study comprises three distinct parts:

data gathering & pre-processing, data conversion and classification, illustrated in Figure 1. It should be noted that this research is only concerned with the classification of malware into distinct family classes and as such no benign samples were used for this research.

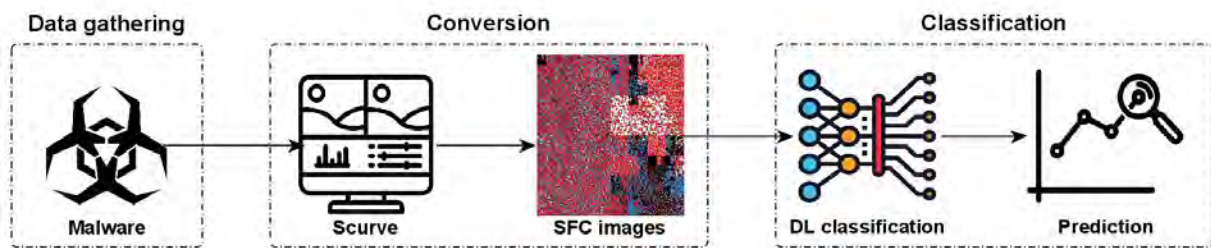


Figure 1: Deep learning malware classification framework architecture

3.1 Data Gathering and Pre-processing

This phase involved gathering and preparing the malware samples for conversion and classification. The samples were collected from the VirusTotal (VT) academic collection (VirusTotal, 2021). The data used to train the deep learning models consisted of 13,091 malware samples obtained during the period January 2018 - December 2019. A further dataset was compiled from January 2020 to July 2021 for validating the models. The validation set was compiled from a more recent timeline to demonstrate the robustness of the models on newer generation variants. It should be noted that due to the lack of samples from some families in more recent repositories, the validation dataset comprised 12,816 samples from 16 out of the 23 families in the training set. Since the samples were not labelled, it was a necessary pre-processing step to cluster the data into their respective family classes. The labels for the samples were obtained through the AVClass labeller tool (Malicia, 2016). This tool parses the VT reports that accompany each sample and predict a class label according to the majority anti-virus detections. This enabled the fast processing of the samples into their distinct family classes.

3.2 Malware Conversion

Space filling curves were chosen as the traversal mapping to convert malware binaries to image format. Space-filling curves are a mathematical construct whose range contains the entire 2-dimensional unit square or more generally an n -dimensional unit hypercube. In this case, the 2-dimensional unit square represents an image of $n \times n$ pixels. SFC's trace a continuous curve through every unit square, i.e., pixel in the image.

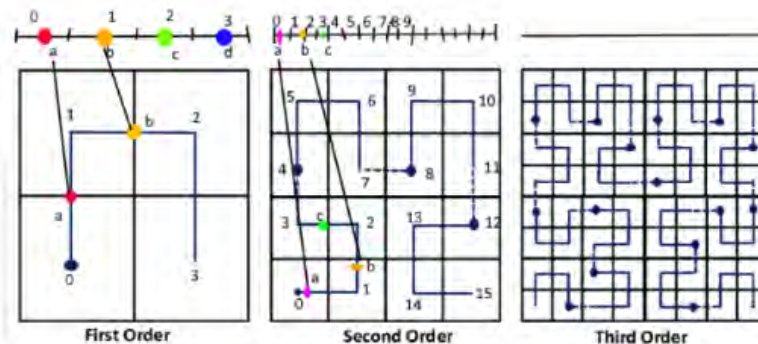


Figure 2: 1-dimension to 2-dimension mapping through Hilbert curve traversal. (a) first order, (b) second order and (c) third order.

Figure 2 depicts the mapping of a 1-dimensional line (representing the sequential malware binary code) to a 2-dimensional curve representation, in this case the Hilbert curve, in first, second and third order iterations. It is evident that closely located points (a and b) map to similar positions on the SFC image. The significance of the SFC traversal mapping in this context is the locality preservation of the data in the resulting images, i.e., similar regions in the malware binary code are grouped together by clusters or regions of distinct colours or textures. In this regard, it has been demonstrated previously by O'Shaughnessy (2019) that malware variants, who by nature share considerable sections of code, map to similar textures within the SFC images. This then allows for the images to be used as representations of the malware for classification purposes.

For the purposes of this research, *scurve*, a Python visualisation library, was used to convert the malware binaries into SFC image format (Cortesi, 2012). The *scurve* library uses a colour-coding scheme, based on data type, that maps the malware binary data to the resulting SFC images. The colour scheme classifies bytes into the following categories: black for 0x00, blue for ASCII text, red for high-value bytes, and white for 0xff. These mappings give the SFC images distinctive texture patterns. The images in Figure 3 show samples from three malware families: Dridex, Emotet and Sfone. If the data type from the malware binary is distinct, it is displayed as non-overlapping sections of a single colour. For example, Figure 3 (a) contains distinct regions of black, denoting binary zeros whereas Figure 3 (b) contains regions of distinct blue, denoting ASCII characters. The rest of the images comprise a mixture of multiple data types, which give the resulting image distinct texture regions. The SFC traversal method chosen for this research was the H-curve implementation, which has been demonstrated previously to possess the most favourable locality preservation properties over other curves. Furthermore, the Hilbert, Z-order and Gray-code curves have previously been explored, whereas there currently exists no work utilising the H-curve traversal.

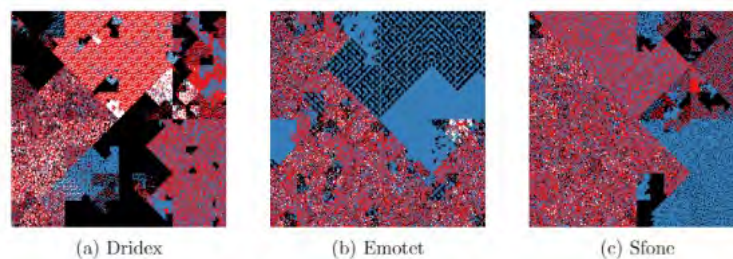


Figure 3: SFC samples in H-curve format

3.3 Classification

The aim of classification in this context is to order the malware samples into their taxonomic groups or family classes, based on shared textural similarities. Convolutional Neural Networks (CNN's) were employed in this research as they had not been explored in similar works previously. This form of classification has been demonstrated as a viable method for image recognition in many different areas. The main classification model was implemented using python Tensorflow (Tensor flow, 2021) and Keras (Keras Team, 2021).

3.3.1 Hyperparameter Auto-tuning

The DL models were constructed through an extensive hyperparameter tuning process using the three Keras auto-tune functions namely, Random Search (RS), Bayesian Optimisation (BO) and Hyperband (HB). RS operates by selecting a random set of HPs and testing the models in an iterative approach, selecting a random set of HPs for each iteration. BO treats the HP set as a regression problem which seeks to find the minimal loss in the minimal amount of processing time. HB provides an extensive search by iteratively sampling combinations of HPs over a defined number of epochs. It selects the optimum set based on the results, which is then used to conduct the final training.

The models trained consisted of a basic model structure as did the base model used while identifying the best HP set for the models. The range for the number of layers that the models were trained on was set between 2 and 10. This was done to keep complexity as low as possible to ensure the training could be carried out within the confines of the available computational power. It was found that 10 layers made it infeasible to execute the entire training process and so it was set as the upper bound for the number of layers in each model. Rectified Linear Unit (ReLU) and Softmax activation functions were utilised in both the model training and the final model. The ReLU function was applied at the convolutional and dense layers of the model, and Softmax was applied to the final output layer to classify each malware sample input into its respective family class.

3.3.2 Model Architecture

In the pursuit of an architecture that is in line with the novel approach identified, the option of a predefined architecture was discarded. The reason being that, while there are several pre-defined CNN architectures that have proven effective in image classification, such as Res-Net50 or VGG16, their complexity and subsequent burden on computational power made them infeasible for the proposed architecture. Instead, a simplified custom architecture was devised, where high classification output performance was achievable, while remaining within the feasible limits of computational power. Furthermore, since the SFC images were represented in a

pixelated form, there was less need for the intensive filtering used on the images as is typical in pre-defined architectures. The models that were constructed for this research required just enough complexity to be effective for classification, while remaining simplistic enough to execute within the limits of the analysis machine, referred to as the "Goldie-locks-zone" (Fort and Scherlis, 2018).

Through the implementation of the models with the auto-tuning process, a network architecture was identified for the final model. This structure was determined through an iterative process of adding a convolutional layer with corresponding max-pooling and dropout layers until the optimum results were achieved. In the final model, discussed in Section 4.2, the outputs of each convolutional layer are connected to a dense layer and a flattening layer, then to the output layer with a softmax activation to classify each input into their respective classes of malware. This approach classifies the one-dimensional representation of the malware binary files using two-dimensional texture patterns of each SFC image as features. The proposed CNN architecture, illustrated in Figure 4, consists of 5 convolutional layers, 7 dropout layers, 3 dense layers and a flattening layer.

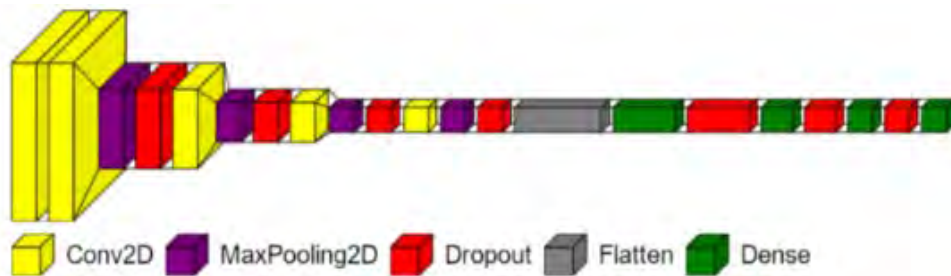


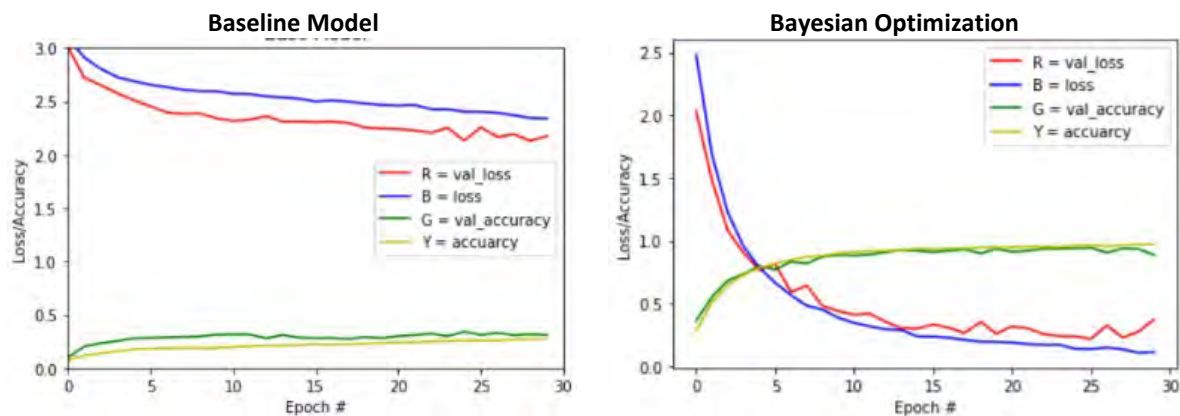
Figure 4: Final model architecture.

4. Analysis & Results

This section discusses the experimental methods devised to optimise the CNN architecture in terms of classification performance and time complexity. Optimal hyper parameters were determined through the auto-tuning functionality of the Keras-Tuner library. The metrics used to determine classification performance were primarily accuracy, precision, recall and F1-score. Plots were used to provide a graphical view of the model's loss and accuracy over the range of epochs used.

4.1 Model Hyperparameter Tuning

A default configuration was devised where no tuning took place, in order to obtain a baseline accuracy to gauge the performance of the auto-tuners. This baseline HP architecture was determined from previous research in image-based classification, such as Le et al. (2018). The baseline and Keras auto-tuners were then applied to the dataset. The loss and accuracy plotted over time (epochs) for each configuration is shown in Figure 5. The x-axis represents the number of epochs, and the y-axis represents loss and accuracy for the training and validation stages. From Figure 5, the baseline training and validation loss are convergent, meaning that the model did not overfit the data. However, the training and validation accuracies are low, with none above 55%. The auto-tuners performed considerably better, as expected. Both the BO and RS compared similarly for training and validation phases, also producing convergent training and validation loss, with accuracy scores above 90% from epoch 10 and saturating at approximately epoch 20.



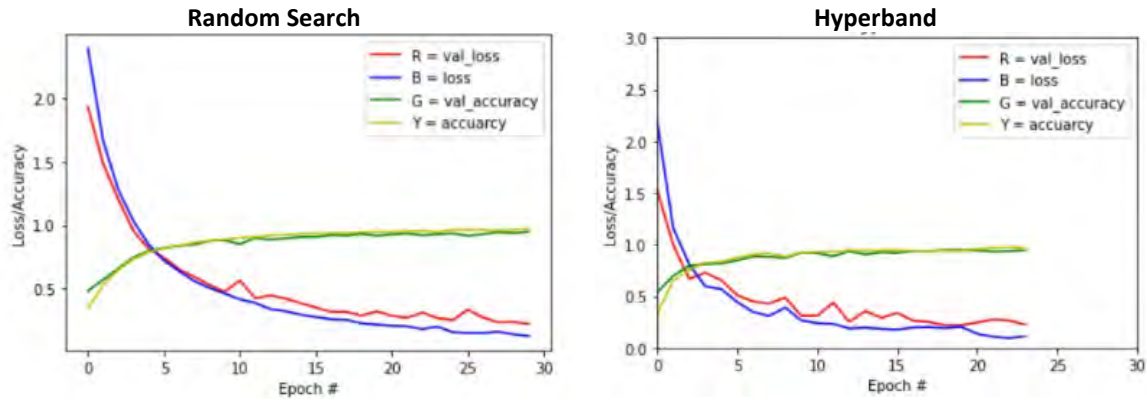


Figure 5: Tuning loss/accuracy performance

The classification performance of the baseline and auto-tuners was next investigated. Table 1 shows the performance and time complexity for each method. The baseline model returned performance metrics of 61%, 48% and 46% for precision, recall and f1-score respectively. This is in line with the outputs from the loss and accuracy plot discussed in Figure 5. The BO tuner performed less favourably than both the RS and HB tuners, with precision recall and F1-scores of 92%, 89% and 89%, which constituted a loss of between 3% and 5%. Furthermore, BO tuning was the slowest of the tuners, constituting almost one second per sample in the training phase. Both RS and HB yielded the same results, with precision, recall and F1-scores of 95%. However, the RS tuner HP set was chosen for the final model as training was almost twice as fast as HB.

Table 1: Model HP training classification performance and time complexity

Tuner Type	Precision %	Recall	F1 Score %	TPS (train)	TPS (val)
Random Search	95	95	95	0.451	0.040
Bayesian Opt.	92	89	89	0.937	0.119
Hyperband	95	95	95	0.823	0.118
Baseline (no tuning)	61	48	46	0.697	0.040

TPS = Time per sample

4.2 Final Optimised Model

The final model was constructed from the optimised HP RS set, coupled with an architecture comprising 5 convolutional layers, 7 dropout layers, 3 dense layers and a flattening layer, obtained through an iterative process, as discussed in section 3.2.2. The number of epochs chosen for the final model was 20, as accuracy stabilised beyond this point. The recorded increased validation accuracy between epoch 18 and 30 was negligible and so an early stopping parameter was set when training the model. When it reached a given value, the gains from further training were deemed negligible and the model stopped training. Further research into the optimum number of epochs would be irrelevant in this case as the dataset is not large enough to warrant a more extensive training process.

4.2.1 Model Validation

If the final deep learning model is to be considered robust, it must be impervious against various perturbations in malware variants as they evolve or mutate over time. Therefore, the final model was evaluated using a validation dataset that was collected from a later timeline, as discussed in Section 3.1. The loss and accuracy plot for the final model on the validation data is shown in Figure 6. The model achieved low training and validation loss values, indicating that the classification error rate was low. Additionally, both loss curves are convergent, which means the model did not overfit the data, so generalised well to the previously unseen data. The model also returned a high classification accuracy for both training and validation, reaching ~95% after epoch 5 and stabilised for the remainder of the experiment. These outputs show that the model was robust in predicting newer generations of family variants.

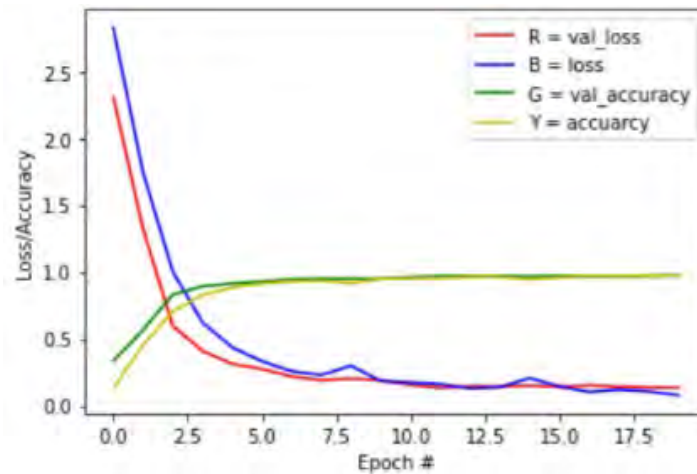


Figure 6: Final model loss and accuracy performance

Table 2 shows the classification performance of the final model for each malware family class. While the model performed well overall, there were some families that achieved low classification rates. For example, the model performed poorly on the Emotet and Scar families. In both cases, precision and accuracy scores were low. Low precision in this case means that when the model predicted a specific class of malware, in many cases it was incorrect. Low recall means there is a high incidence where the model incorrectly predicted other malware families as a particular class. However, the dataset compiled for this research is imbalanced, due to the lack of availability of some family samples. From Table 2 the sample count, both Emotet and Scar contained low numbers of samples (44 and 91 respectively), in which case there may have been insufficient features to train the models on, leading to poor classification performance.

The class imbalance issues were compounded by the evolving nature of the malware. To provide illustrative examples, Figure 7 shows Emotet samples taken from the training (a-c) and validation (d-f) datasets. It is evident that the texture patterns are different across all example SFC images. Emotet has undergone many changes and evolutions since its inception in 2014, where Emotet's developers have changed its functionality and increased obfuscation features in order to evade detection.

Table 2: Performance of the final model on the validation dataset

Malware family	Precision %	Recall %	F1 Score %	Sample Count	% Of sample count
Allaple	81	97	88	264	1.92
Berbew	99	71	82	4146	35.35
Dinwod	98	96	97	911	7.1
Dorkbot	56	95	71	130	1
Emotet	7	55	13	44	0.34
Fsysna	88	72	79	188	1.47
Hematite	94	99	97	670	5.23
Oberal	33	97	49	101	0.79
Picsys	94	99	97	295	2.3
Salgorea	98	93	95	1235	9.64
Scar	16	19	17	91	0.71
Sfone	99	98	99	2719	21.21
Socks	79	100	88	30	0.23
Sytro	96	87	91	991	7.73
Vilisel	93	100	97	70	0.55
Vobfus	94	98	96	931	7.26
Weighted avg.	95	86	90		

This results in the code structures differing greatly as the variants evolve. This produces disparate SFC images as displayed in Figure 7 which in turn had a detrimental effect on the prediction capabilities of the models.

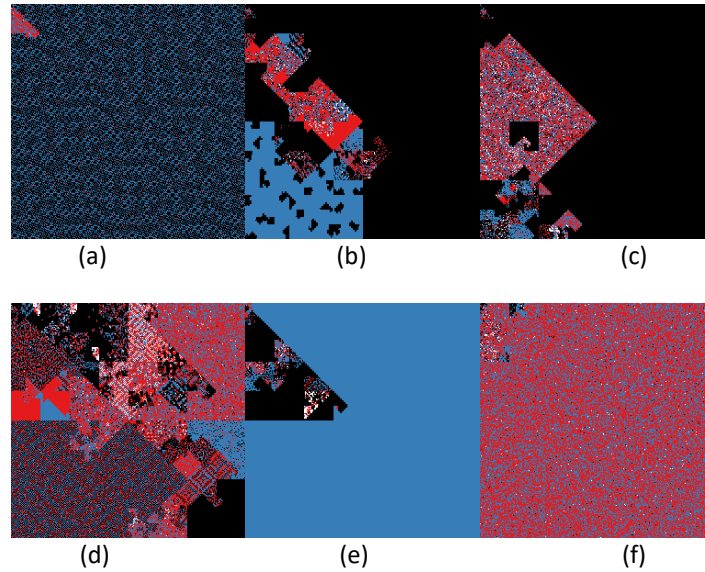


Figure 7: Emotet samples taken from the training (a-c) and validation (d-e) datasets

4.2.2 Time Complexity Optimisation

The initial conversion process produced SFC images of 256 x 256 pixels as this is the default setting in the scurve library. However, the average time complexity per sample to train the model was 8.8 seconds, which can be considered infeasible for large datasets. The optimisation of time complexity through smaller dimension images was thus considered. The size parameter in the scurve library was tested at various dimensions and it was found that training time was greatly reduced to 1.1s per sample, when the SFC images were converted to 32 x 32 pixels, resulting in an accuracy loss of just 1.1%. This allowed the training and subsequent validation testing to be conducted within a reasonable timeframe, which enhances the scalability of the proposed method. It should be noted that the final tuned model achieved moderately faster time complexities of 0.925s and 0.041s per sample for training and validation, respectively.

4.2.3 Comparison with previous benchmark method

To test the feasibility of the proposed deep learning method for image-based malware classification, a comparative analysis was conducted with the previously proposed method by O'Shaughnessy (2019), discussed in Section 2. This method can be considered a benchmark as it was the first documented work using space-filling curves for malware classification. The best performing model, KNN-HOG, was chosen for comparative purposes. The KNN-HOG model was trained and evaluated on the same H-curve SFC training and validation sets as the proposed deep learning method. Table 3 shows the performance results for training and validation for both methods. From the table, KNN-HOG outperformed the proposed method marginally in the training phase, with precision, recall and accuracy scores of 97%. However, in the validation set, the deep learning model performed considerably better in all three performance metrics. The results show that the proposed deep learning model is more robust than the benchmark model, which demonstrates its feasibility as a viable method for image-based malware classification.

Table 3: Performance comparison for KNN-HOG and DL models on training and validation data

Model	Training			Validation		
	Precision %	Recall %	F1-score %	Precision %	Recall %	F1-score %
KNN-HOG	97	97	97	83	63	63
DL	95	95	95	95	86	90

5. Discussion

The application of deep learning to classify SFC malware representations had not previously been explored, which prompted the main motivation of the study. Furthermore, previous research focused on gathering features for classification through invasive or complex malware analyses methods. The framework presented in this paper provides a method for malware classification that negates the need for intensive feature generation or invasive analysis techniques. By utilising convolutional neural networks, feature extraction and selection is not necessary as the models identify different levels of image representations by learning the basic textures in

the first layers and evolving to learn features of the image in the deeper layers. The CNN architecture can be considered shallow in that it consists of just 5 convolutional layers, which simplifies the model's complexity for more efficient computation, enhancing scalability for larger datasets. Additionally, the final model returned high prediction rates on a validation set of variants from a more recent timeline, demonstrating its robustness against evolved malware family variants. Finally, in order to evaluate the viability of the proposed method, a comparative analysis was conducted with the final model and the benchmark method by O'Shaughnessy (2019). The results show the final model significantly outperformed the benchmark model on the validation dataset, proving its viability.

6. Conclusions and Further Work

The main aim of this research was to produce a malware classification solution that was robust and scalable to overcome current limitations. The findings presented here are significant in that they demonstrate the feasibility of proposed deep learning framework as a robust, scalable method for image-based malware classification. Limitations exist, including the size of the dataset and the format of the malware. CNNs have been shown to perform best on larger datasets. There is scope to increase the malware dataset to improve the model's performance. The malware format chosen to investigate was the Windows Portable Executable format, which represents the format used in the majority of malware produced. Again, research could be expanded to include other formats such as malicious PDF or Word Macros.

References

- AV-Test, 2021. Malware Statistics & Trends Report | AV-TEST [WWW Document]. URL [/en/statistics/malware/](#) (accessed 2.18.21).
- Conti, G., Bratus, S., Shubina, A., Lichtenberg, A., Ragsdale, R., Perez-Aleman, R., Sangster, B., Supan, M., 2010a. A Visual Study of Primitive Binary Fragment Types 17.
- Conti, G., Bratus, S., Shubina, A., Sangster, B., Ragsdale, R., Supan, M., Lichtenberg, A., Perez-Aleman, R., 2010b. Automated mapping of large binary objects using primitive fragment type classification. *Digital Investigation* 7, S3–S12. <https://doi.org/10.1016/j.diin.2010.05.002>
- Conti, G., Dean, E., Sinda, M., Sangster, B., 2008. Visual Reverse Engineering of Binary and Data Files, in: Goodall, J.R., Conti, G., Ma, K.-L. (Eds.), *Visualization for Computer Security*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1–17. https://doi.org/10.1007/978-3-540-85933-8_1
- Cortesi, A., 2021. cortesi/scurve, <https://github.com/cortesi/scurve>.
- Fort, S., Scherlis, A., 2018. The Goldilocks zone: Towards better understanding of neural network loss landscapes. *arXiv:1807.02581 [cs, stat]*.
- Fukushima, Y., Sakai, A., Hori, Y., Sakurai, K., 2010. A behavior based malware detection scheme for avoiding false positive, in: 2010 6th IEEE Workshop on Secure Network Protocols. Presented at the 2010 6th IEEE Workshop on Secure Network Protocols, pp. 79–84. <https://doi.org/10.1109/NPSEC.2010.5634444>
- Institute, M.L.@ I.S., 2021. AVClass and AVClass2.
- Kaggle, 2015. Microsoft Malware Classification Challenge (BIG 2015) [WWW Document]. URL <https://kaggle.com/c/malware-classification> (accessed 3.3.21).
- Keras Team, 2021. Keras documentation: Developer guides [WWW Document]. URL <https://keras.io/guides/> (accessed 6.21.21).
- Khalilian, A., Nourazar, A., Vahidi-Asl, M., Haghighi, H., 2018. G3MD: Mining frequent opcode sub-graphs for metamorphic malware detection of existing families. *Expert Systems with Applications* 112, 15–33. <https://doi.org/10.1016/j.eswa.2018.06.012>
- Kolter, J.Z., Maloof, M.A., 2004. Learning to detect malicious executables in the wild, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04. Association for Computing Machinery, New York, NY, USA, pp. 470–478. <https://doi.org/10.1145/1014052.1014105>
- Le, Q., 2018. Deep learning at the shallow end: Malware classification for non-domain experts. *Digital Investigation* 9.
- Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S., 2011. Malware images: visualization and automatic classification, in: Proceedings of the 8th International Symposium on Visualization for Cyber Security - VizSec '11. Presented at the 8th International Symposium, ACM Press, Pittsburgh, Pennsylvania, pp. 1–7. <https://doi.org/10.1145/2016904.2016908>
- Nataraj, Lakshmanan, Yegneswaran, V., Porras, P., Zhang, J., 2011. A comparative assessment of malware classification using binary texture analysis and dynamic analysis, in: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence - AISec '11. Presented at the 4th ACM workshop, ACM Press, Chicago, Illinois, USA, p. 21. <https://doi.org/10.1145/2046684.2046689>
- O'Shaughnessy, S., 2019. Image-based Malware Classification: A Space Filling Curve Approach, in: 2019 IEEE Symposium on Visualization for Cyber Security (VizSec). Presented at the 2019 IEEE Symposium on Visualization for Cyber Security (VizSec), IEEE, Vancouver, BC, Canada, pp. 1–10. <https://doi.org/10.1109/VizSec48167.2019.9161583>
- Tensor flow, 2021. Libraries & extensions | TensorFlow [WWW Document]. URL <https://www.tensorflow.org/resources/libraries-extensions> (accessed 4.21.21).

- Vasan, D., Alazab, M., Wassan, S., Safaei, B., Zheng, Q., 2020. Image-Based malware classification using ensemble of CNN architectures (IMCEC). *Computers & Security* 92, 101748. <https://doi.org/10.1016/j.cose.2020.101748>
- VirusTotal, 2021. VirusTotal – Learning resources. URL <https://www.virustotal.com/learn/> (accessed 8.28.21).
- Wagner, M., Fischer, F., Luh, R., Haberson, A., Rind, A., Keim, D.A., Aigner, W., 2015. A Survey of Visualization Systems for Malware Analysis. Presented at the Eurographics Conference on Visualization (EuroVis), pp. 105–125. <https://doi.org/10.2312/eurovisstar.20151114>

Integrating Democratic Cybersecurity: Empowerment of Traditional Law Enforcement and Democratic Public Safety

Michael Losavio, Jeffrey C. Sun, Sharon Kerrick, Adel Elmaghraby, Cheryl Purdy and Clay Johnson

University of Louisville, Louisville, Kentucky, USA

Michael.Losavio@louisville.edu

Jeffrey.Sun@louisville.edu

Sharon.Kerrick@louisville.edu

Adel@louisville.edu

Cheryl.Purdy@louisville.edu

Clay.Johnson@louisville.edu

Abstract: The expansion of pervasive and ubiquitous computing, especially with the advancement of the Internet of Things and the Smart City concept, extend the novel means of criminality and its investigation. We argue that current forms of investigation and discovery are not sufficient to limit injuries onto persons and communities. Nonetheless, cybersecurity approaches within criminal justice, criminology, and workforce development – together – offer models that significantly benefit efforts to address public cybersecurity harms, yet they have been largely overlooked. This paper draws on an interdisciplinary lens to address cybersecurity, including criminal justice and workforce development integration and employing empowerment theory. Applying empowerment theory, this presentation demonstrates the effects from integrating cybersecurity and forensic practices into traditional law enforcement. The effects are positive as public safety will be needed to provide public safety and security in our hybrid technical world. Thus, this paper illustrates how we must, in essence, “democratize” cybersecurity through its distributed availability. We present means to achieve this and results from efforts to promote this integration through several coordinated, yet differently targeted programs at one research university.

Keywords: democratic, distributed, cybersecurity, law enforcement, public safety

1. Introduction

The expansion of pervasive and ubiquitous computing, especially with the advancement of the Internet of Things and the Smart City concept, extend the novel means of criminality and its investigation (Chow, Losavio, Joshua, and Koltay 2018). Indeed, higher education has tackled the privacy, breach, and harms associated with data analytics and cyber for many years (Sun 2014), but its solutions are relegated to a specialized group so operations continue in spite of the high degrees of harm and continuous threats. We argue that current forms of cyber investigation and discovery are not sufficient to limit injuries onto persons and communities.

Traditional public safety relies on distributed efforts of citizens, police, courts, and corrections. By contrast, cybersecurity efforts have tended towards limited points of security via IT technical systems and guardians. Our societal solutions focus on narrow expertise and limited actors. In many nation-states cyber crisis employs a public-private partnership to address attacks, but often one entity takes the lead. For instance, Czech and Estonian responses centers around a “network administrative organization” approach where an outside entity manages the network activities with enforcement power (Boeke 2018). Denmark operates off a “lead-agency” model, and Netherlands established a “participant-governed network” to better link the private organizations with the government agencies (Boeke 2018).

This paper applies empowerment theory. While the term empowerment is somewhat of a hackneyed saying and captures many constructs, some of which are not empirically sound, here, the concept of empowerment theory is more than simply encouraging and motivating others (Wilkinson, 1998). Broadly speaking, empowerment theory connects processes and structures that individuals and groups employ to reach common goals (Perkins & Zimmerman, 1995). It is conceptualized as a process and outcome theory; macro and micro theory (with consideration of the meso range); and organizational, intrapersonal, and individual levels (Amor, Xanthopoulou, Calvo, & Vázquez, 2021; Perkins & Zimmerman, 1995). For purposes of this article, we adopt Leong, et al.’s (2015) empowerment theory. It is an applied crisis response version of that considers the structural, psychological, and resource empowerment settings. This model connects relationships between empowerment settings, in which structural and psychological empowerment are connected with the attainment of shared identification of the activity or outcome, structural and resource empowerment are connected with the attainment of collective

participation, and resource and psychological empowerment are connected with the attainment of collaborative control toward the outcome. Given its nesting in information and communication technology (ICT) and application in a crisis management setting, its use as the type of empowerment theory for cybersecurity seems appropriate.

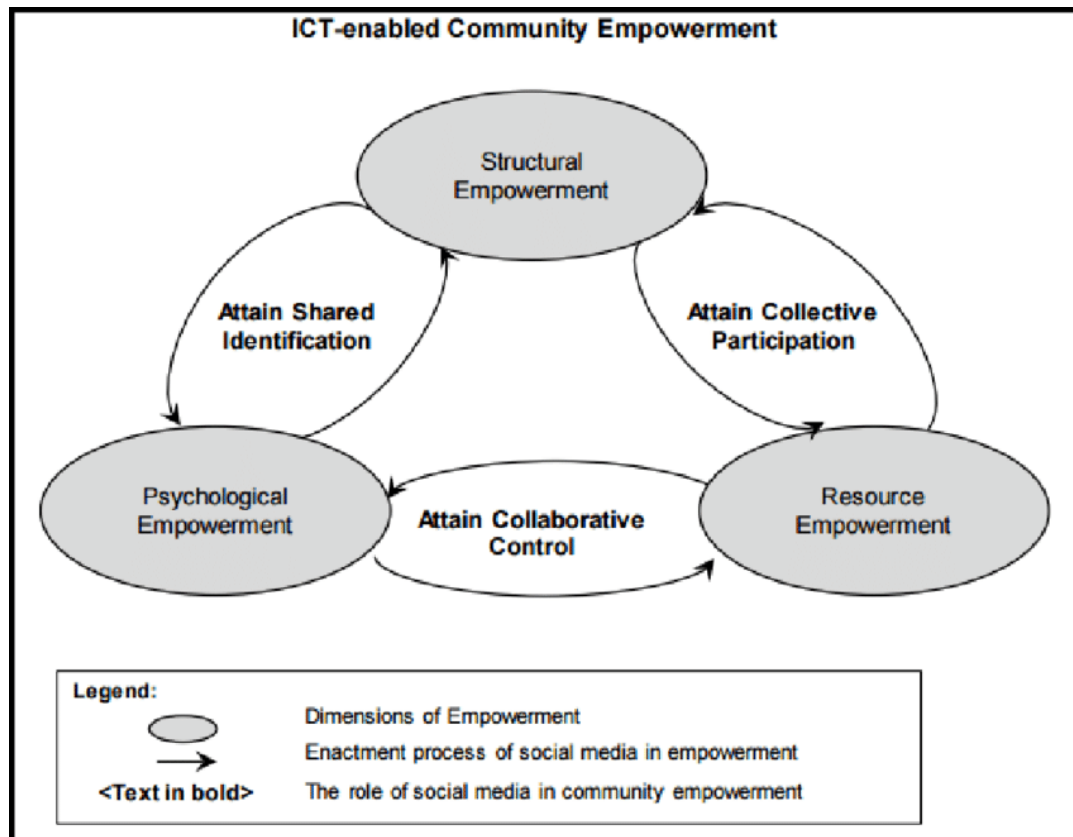


Figure 1: Leong, et al.'s (2015) information and communication technology enabled empowerment theory

Drawing on empowerment theory, the authors demonstrate, through multiple-related cyber training programs, that the rapidly shifting information technology environment make it is necessary to expand public and citizen engagement in cybersecurity. This approach is to, in effect, take principles of democratic knowledge and governance and extend them into the realm of security for cyber-resources by empowering people around us. The goal is empower through structural, psychological, and resource empowerment settings to build a workforce in cyber but drawing not simply on development of cyber experts in computer science, engineering, or information technology. Instead, we develop a generalized and wide scope of cyber workforce akin to "leadership" allowing many to enter this space, not selected few who reside in a tech-tower. Stated another way, this approach demonstrates how we must empower small organizations, from small businesses to non-governmental organizations (NGOs) to local law enforcement agencies to educational institutions, as well as the citizens themselves, to engage in cybersecurity.

This application of the empowerment theory presents solutions to the principles in criminology for effectively reducing crime, such as the Routine Activity Theory and the Opportunity Theory. Routine Activity Theory and the Opportunity Theory opine that crimes occur during vulnerable stages through opportunity seeking behaviors of criminals. Through empowerment theory, we envision opportunity reductions and closures in cyber gaps. Accordingly, these implementations to close these gaps and design empower-thriving settings (among structural, psychological, and resource empowerment settings) cyber breaches will be reduced. That is, a distributed cyber workforce development at the micro-level, which may be expanded across domains of multiple professions, much like leadership training and development, will protect people by reducing the number of vulnerable targets subject to attack. At the same time, by expanding available "guardians" as to reduce and deter those intending to participate in wrongdoing, we will have enhanced our cyber public safety.

It is essential for cyber-safety to anticipate future risks and vulnerabilities to systems. To do so, it is important to ask whether there are constantly changing issues as to greater vulnerabilities or greater disruptions, and whether there can be effective technical protections. For physical capabilities, our policymakers, organizational leaders, and educational institutions have a sense of how to avoid many risks and protect people from vulnerabilities at concerts, sporting events, community gatherings, and festivals. The challenge is associated with cyber attacks when they involve computers, networks, and the cloud, but they also involve cell phones, watches, cars, televisions, and even refrigerators. Although this paper addresses a regional effort, this concern of cyber breaches is, on the broader note, to be considered as a national/international issue threatening all of us and nearly all things. Accordingly, we are charged (and could be empowered) to look at potential solution sets and ask whether various activities or interventions are scalable by size of nations and whether they are portable to varied nations (or even areas/regions).

Globalization has led to integration across many dimensions. It will be important that as many as possible can be protected, even as the ubiquity of computing will grow with the Internet of Things, the Smart City, Big Data, and universal and pervasive computing generally, as seen in Figure 2.

And with tomorrow's IOT, Smart City, Ubiquitous
networking, Big Data...
(slightly exaggerated...)



Figure 2: The Integrated Domains of Public Safety, Cybersecurity and Digital Forensics

Under the democratic accountability implicit in Western constitutions, law enforcement and public safety are vital to the liberties and protections of citizens. Together, they may assure better community protection and engagement under the rule of law. Standing alone, however, neither is sufficient. One example of a close engagement between law enforcement and the people is community policing as an approach. Community policing involves assigning officers to certain areas so they become familiar with the local citizens in their assigned area. If applied to cyber security, we must consider several questions including:

1. What is a community for cyber security?
2. How do we embrace and empower communities to understand and observe the community norms or notice outlying coded behaviors?
3. What alternatives might we have to cyber community policing?

Several scholars suggest that we resort to the provisions under the U.S. Constitution to issue letters of marque and reprisal against malicious entities for counter-cyberattacks (Ayers, 2021). However, the question then becomes: would that really work, or would it just open up the Pandora's cyber-box of furies to haunt us? Thus, we posit that an integrated system, which is based on community empowerment of public safety, is needed for cybersecurity.

2. Case Study-Ransomware and a Distributed Model Protection

Cybercrime and extortion are moving down to the consumer level, hitting small businesses as well as large enterprises. Styled "Ransomware," this attack consists of malware, like a computer virus, that has been downloaded to your system and encrypts all of your files, then shows a ransom note stating that you will get the decryption key if you pay the amount demanded. If the attacker demands bitcoin or similar digital tokens, it is difficult, if not impossible, to trace. If you do not pay the "ransom", at best you may lose use of those files; at worst, critical, confidential documents will be posted on the Internet. Major businesses have been hit, and many have

paid millions to avoid critical service disruptions, ranging from energy infrastructure (e.g., fuel pipelines) to medical systems (e.g., hospitals).

Cybercrime is a serious risk for small businesses, organizations, and families, coming from an explosion in ransomware attacks tied to its profitability. For example, Forbes Magazine (Shankar 2021) noted that ransomware operators are targeting law firms, which are usually small businesses that are data and information intensive. The confidentiality of information that law firms possess is vital, and there is a professional obligation subject to sanctions for its violation. The motivation is, as always, money, whether it be directly or indirectly. Consider the value of confidential data from corporate entities, it is ideal for investing ahead of public announcements about data breaches such as sensitive information like individual tax returns or breaches making them perfect for identity theft fraud. Forbes cited the American Bar Association's *TechReport 2020: Cybersecurity*, which noted significant problems with cybersecurity across law firms and the accompanying potential ethical risks for lawyers (Loughane 2020). The risks of losing client data to the opposing party or a third party wishing to disseminate the information beyond litigants present real threats in terms of financial, psychological, and possibly physical means (as the disclosures often evoke long-term physical reactions).

The Cybersecurity & Infrastructure Security Agency ("CISA") of the U.S. Department of Homeland Security, in an effort to help guide those without their own IT security units, built online guidance and resources to help the distributed operations for our society (Cybersecurity and Infrastructure Security Agency 2020). The CISA Ransomware Guide, available for free download, provides a resource on the "best practices" to prevent attacks and a checklist to respond and recover, if and when, an attack occurs (Cybersecurity and Infrastructure Security Agency 2021). Because the pessimists in cybersecurity opine that it is not a matter of "if," but only "when," a cybersecurity breach will happen, response and recovery planning is essential.

The CISA Ransomware Guide opens with a review of the best practices for the prevention of a cyber attack, beginning with the preservation of protected off-line backups through regular maintenance for vulnerabilities. It covers the benefits and limitations of standard cybersecurity systems and notices the particular role played by the human condition in the success of employee-targeted phishing attacks. The self-help tools offered by CISA then provide a checklist to review your preparation for threat/risk prevention discussions with third parties that may have access to target systems.

This guide still requires due diligence on the part of the citizen-user. This due diligence can range from preparatory talks with IT specialists about future emergencies to consultation with professional liability insurers on both the coverage they offer and the support they can provide in the event of an attack. And, yet again, the training of employees and users on the proper practices for cyber-safety and security is essential. The technical is simply not enough, and by reinforcing this across an enterprise, security is enhanced, hopefully enough to prevent compromise.

3. Distributed Cybersecurity, Local Law Enforcement, and Democratization of Cybersecurity

The democratization of technology and the democratization of knowledge are sometimes used to address the process by which both become more available to more and more people in society. This ranges from the open-source movement, open-source meaning "software in which the source code is freely available for others to view, amend, or adapt" (Poynder 2001), to the ease of use of proprietary devices. It supports inclusion, opportunity, and accountability.

The University of Louisville has focused on issues relating to distributed democratization of cybersecurity, with analysis and position papers regarding the importance of cybersecurity as part of overall governance in a world of ubiquitous computing (Keeling, Shutt, Losavio 2011), (Elmaghraby, Losavio 2017), (Keeling, Losavio 2009) (Losavio, Elmaghraby 2021) (Chow, et al 2018).

These papers led to support from the National Security Agency of the U.S. Department of Defense in 2017 via their Cyber Workforce Development Program. This project was directed by Dr. Adel Elmaghraby, Director of Research and Innovation for the University of Louisville's Digital Transformation Center. The concepts behind this program are based on the changing context of cybersecurity criminal conduct in the world. It grew in

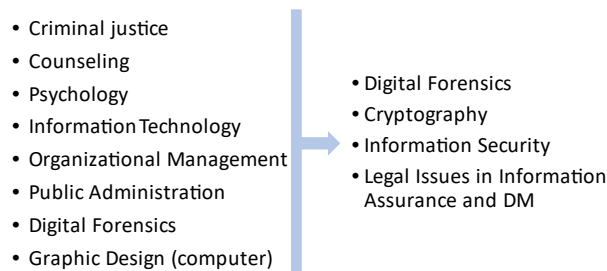
importance because of the ubiquity of computing in the modern world, so cyber attacks could arise from many sources, including nanny cams, emails, text messages, and social media applications.

A decade earlier, the historical context would indicate some, but not much, overlap between the domains of cybersecurity and digital forensics (i.e., the discipline of finding and examining electronic evidence). But such connections were not initially in the works or fully conceived by many of our counterparts. The lack of overlap is clearly changing, such that we must examine the relationships of cybersecurity and digital forensics within the overall structure of public safety, both now and in anticipation of future issues arising from the pervasiveness of computer systems, from the Internet of Things to the Smart City, as shown in Figure 2, above.

To address this cyber readiness, we test various systems relating to local inclination – local needs in defense, local needs in healthcare, local needs of educational institutions, and local needs law enforcement – as they relate to cybersecurity. We examine both the inclination and capability – of defense workers, healthcare workers, educators, and state and local law enforcement to address cybersecurity and cybercrime issues.

This examination of potential online cybersecurity training programs began with five modules for local law enforcement agencies; the program had 111 learners with a waiting list of approximately 100. At least 18 law enforcement occupations, ranging from detectives to criminal intelligence analysts, were present in this first cohort. Officers working in Crimes Against Children and Professional Standards units were in attendance at the first cohort, as were law enforcement personnel from 17 different states, ranging from Alabama to Wyoming. Members of the learning cohort had backgrounds from a surprising range of disciplines, including psychology, organizational management, counseling, criminal justice and public administration, as shown in Figure 3.

Background Areas vs. Target Areas



24

Figure 3: Learner Backgrounds

The five modules covering cybercrimes were law and practice, infrastructure technology, network security, information security, and a review of computer forensics.

None of the learners were compensated beyond the courses themselves and, to the best knowledge, worked on these courses in addition to their regular work requirements or academic obligations in other programs. Learners began to drop out of the courses, with several noting a conflict with work. In one instance, a homicide detective notified via email that he would be late getting her/his work completed due to a new homicide investigation, only to send another email the following week stating that s/he would have to drop the course entirely due to a *second* homicide investigation that had opened. This became an issue throughout the course, shown here in Table 1.

September 2017 until June 2018

Table 1 - Initial Data on Participation in Introductory Courses

Topic	Enrollment	Completion
Cyber Crimes: Law and Practice	111	completed
Infrastructure Technology	67	completed
Introduction To Network Security	59	completed
Introduction To Information Security	47	completed
Computer Forensics	32	completed

25

Expansion of Cyber Security Training into Other Domains

This cybersecurity training for state and local law enforcement was followed by a series of similar, broad-based cybersecurity trainings in other domains. Those projects are:

1. U.S. Department of Defense C4 Training Funding;
2. U.S. National Security Agency CAE Cybersecurity for Healthcare Industry; and
3. U.S. Department of Homeland Security Cybersecurity Training for Law Enforcement Funding.

3.1 U.S. Department of Defense C4 Training

The U.S. Department of Defense C4 project is a cybersecurity workforce development project via the State government of the Commonwealth of Kentucky. It is directed by Dr. Jeffrey C. Sun, professor of higher education and law and associate dean of the College of Education and Human Development for the University of Louisville. Under this project, the Commonwealth of Kentucky requested a \$2 million grant to develop cybersecurity education programs and continue efforts in workforce development for transitioning service members, veterans, and dependents. This grant is in Phase III to diversify Kentucky's defense sector, in alignment with the Economic Adjustment Assistance for State Governments Program. During the period of the grant, funding provides a start-up seed program for developing cyber talent. It is a three-phased operation beginning in October 2019 and ending in March 2022, followed by a transition from a grant-based project to an independent university program. The objectives of the project are to:

- Enable functions to shape, set and administer a world-class cybersecurity program for military-connected participants;
- Execute unique research-based cybersecurity educational pathways;
- Empower employment search success; and
- Grow, scale, and sustain program outcomes.

If these objectives are met, the result will be attainable increased cybersecurity opportunities for Kentucky generally, and for those wishing to work in these fields specifically. Beginning with gap analysis, this project will continue through training cyber experiences, including in the Cyber Range for practical problem-solving in cybersecurity, as detailed in Figure 4:

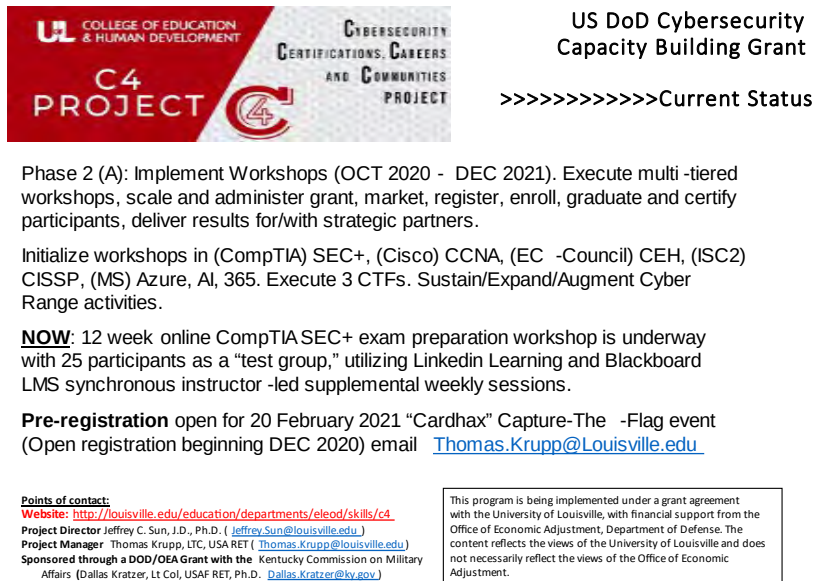


Figure 4: DOD C4 Project Overview

The multiple phases of this project begin with recruitment and extend through the initialization and implementation of workshops and standard credentials, particularly through CompTIA. Information is distributed detailing these objectives and points of contact on the project, as shown in Figure 5:

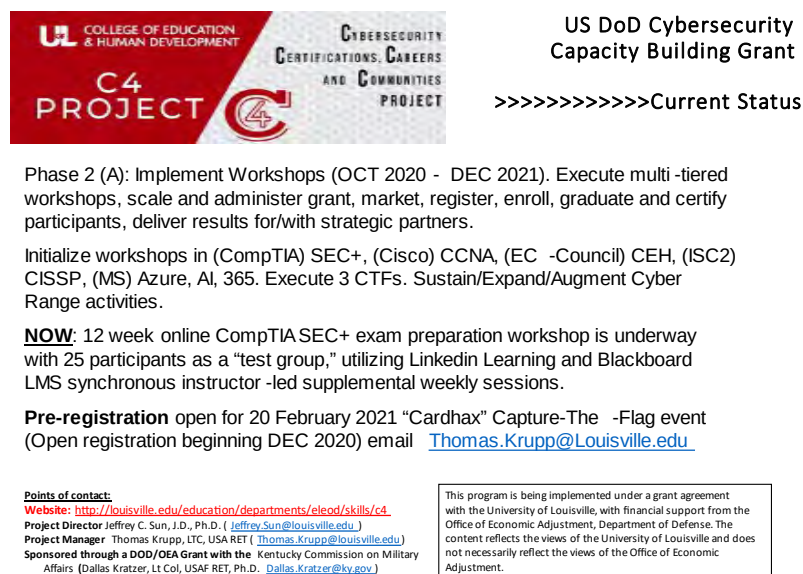


Figure 5: Workshops and Credentials

3.2 National Centers of Academic Excellence (NCAE) Cybersecurity in Healthcare

Concurrently, a project relating to cybersecurity for the *healthcare industry* was implemented under the direction of Dr. Sharon Kerrick, assistant Vice President for Digital Transformation. This project focused on cybersecurity skills within the healthcare industry, a primary target of cyberattacks via ransomware. For instance, in November 2020, the Vermont National Guard were mobilized to aid the University of Vermont Medical Center in recovering systems after a hacking incident (Scott, 2020). In another instance, a wrongful death action has been filed against a hospital for failing to protect against a ransomware attack¹ that led to the death of an infant (Collier 2021).

¹ The party alleged that the hospital failed to inform the mother of the attack prior to or when she arrived to deliver her baby, which resulted in diminished care and no tests being administered to see that the umbilical cord was wrapped around her baby's neck, resulting in severe brain damage and eventual death.

This project tests the use of online, mentored training in three levels of depth in the discipline: Explorer, Practitioner, and Professional. The Explorer module begins with IT basics and security principles through policy and legal /ethical foundations, and it begins to introduce issues relating to coding and artificial intelligence data mining. The Practitioner module dives deeper into information and network security, forensics, cryptography and cyber threat hunting, intermediate data mining, and legal/regulatory issues within the healthcare industry via HIPAA.

Lastly, the Professional module moves further into database security, cloud security, advanced data mining, and Deep Learning/AI, among other topics. These are followed by industry certifications that establish the skills of the learner for current and future employers. This is shown in Figure 6:



Figure 6: NCAE Cybersecurity in Healthcare Programmatic Areas

3.3 The U.S. Department of Homeland Security Law Enforcement Cybersecurity Project

The U.S. Department of Homeland Security 2022 Law Enforcement Project follows the earlier NCAE-funded effort in 2017. The 2022 Law Enforcement Project follows the same general topic structure with enhanced use of online learning systems.

A comprehensive analysis by the Police Executive Research Forum (“PERF”) found local police agencies are still in the early stages of developing comprehensive strategies for responding to and preventing cybercrime. Additional work needs to be done to ensure that every police officer will know how to respond to a cybercrime call; detectives will understand the avenues that are available for investigating these crimes; and local, state, and federal authorities will have an effective system for sharing information, connecting cases committed by the same offenders, and coordinating investigations. It is critical for the police to have the skills to respond to and investigate cybercrime cases as effectively as they handle other types of cases. Communities will expect their local police to be familiar with cybercrime issues, cybercrime prevention, and investigation. Further, community education on cybersecurity was deemed vital. Police departments often offer education programs to teach people about common cyber-threats and the basics about how they can stay safe online.

Sociological, criminological, and criminal justice/law enforcement models support means of programmatic use of law enforcement for cybersecurity. In structuring public safety solutions to cybersecurity, especially those engaging law enforcement, the rules for our system of justice must be incorporated into the solution.

Supported first by the National Security Agency and then by the U.S. Department of Homeland Security’s Science and Technology Directorate, the online training program should increase the pool of cybersecurity professionals in multiple domains by identifying, recruiting, and training practitioners and students in law enforcement and public safety disciplines, including police, probation and parole, military, and other public safety areas. The project does this through a multi-tiered training and education program to develop cybersecurity and forensic skills at the undergraduate and graduate levels. This will include, at a minimum:

- a set of fundamental courses on computer and network operations and security for those in need of basic computer and network operations and security skills; and

- a subsequent series of intensive online courses in advanced cybersecurity techniques and issues for those who have shown exceptional skills in operations and security.

These teaching programs were developed through collaboration by the Department of Criminal Justice; the Department of Educational Leadership, Evaluation and Organizational Development; the Department of Computer Information Systems of the Business College; and the Department of Computer Science and Engineering. We hope cybersecurity practices can be deployed in the community and used both to expand civil cybersecurity protection and to provide skills for the use of law enforcement deterrence and incapacitation of offenders.

This project integrates traditional public safety personnel into the cyber-safety regime. This integration will provide a broad distribution of cybersecurity skills, advice, and counsel across a community which may, in turn, serve to better harden citizen targets—whether that be the elderly, families, or small businesses—against attacks. These attacks may be directly against these victims. Or, they may seek to attack others, who may be more lucrative and vulnerable targets. In turn, these public safety professionals will have cybersecurity skills they can use once they leave public service life and move into other domains that need their skills for cybersecurity (perhaps, corporate consulting groups). Detailing their ability to cross into corporate and other settings, we outline the outcomes of these learners. By participating in the program, learners will:

1. Increase the pool of cybersecurity professionals in multiple domains by identifying, recruiting, and training practitioners and students in law enforcement and public safety disciplines;
2. Integrate criminal justice systems and practices into cybersecurity; and
3. Advise local law enforcement, which is at ground zero for cyber-criminality and its victims, on effective ways to respond, advise, and arrest attackers.

The training, which began in 2021, built on a needs assessment by Professor Jason Gainous that found, *inter alia*, that:

1. Most agencies both internally investigate cybercrime incidents and report instances of cybercrime to other reporting agencies;
2. Generally speaking, agencies are lacking in both general and specialized training as part of their basic training, or as part of in-service operation, and do not have dedicated cybersecurity and cybercrime units that have staff with a specialized focus; and
3. Agencies are short both on offering community training and on resources for community training.

The training cohorts began with a total number recruited of 174, but only 171 continued. Regarding the rank of those recruited, there is an almost even split for most common, with police detectives/investigators (n=34, 19.5%) just slightly outnumbering patrol officers/troopers (n=33, 18.9%). The two most common ranks were followed closely by police lieutenants (n=27, 15.5%). Police captains, sergeants, and non-sworn personnel were equally common (n=15, 8.6%) and were slightly more common than police chiefs/assistant chiefs (n=13, 7.4%). The least common among the ranks were majors (n=4, 2.3%), deputies (n=3, 1.7%), corporals (n=2, 1.1%), commanders (n=1, 0.6%), and sheriffs (n=1, 0.6%). Other titles/ranks registered included agents/special agents (n=4, 2.3%), instructors/directors (n=2, 1.1%), and miscellaneous military designations (n=2, 1.1%).

As to the states, cities, and departments participating, there are 32 states participating, including Washington D.C. (see full list below), and registration interest was received from officers in Nigeria and Beirut as well.

Across the United States, individuals from police departments, sheriff's offices, and other governmental departments have registered for the course using 117 unique addresses and just as many zip codes. This suggests that our course has permeated departments and has been spread among divisions within the same city and department. The U.S. states represented include Pennsylvania, Indiana, Massachusetts, Michigan, New York, Kentucky, Illinois, Tennessee, Florida, North Carolina, Arkansas, South Carolina, Vermont, Mississippi, Missouri, Ohio, Alabama, Washington, South Dakota, Connecticut, California, Georgia, Oregon, Virginia, North Dakota, Iowa, Kansas, Colorado, Nebraska, Minnesota, Arizona, and Washington D.C.

This data demonstrate both an existing need for the skills among local law enforcement and an interest in training to provide the skills across multiple dimensions. The variety of law enforcement departments in the multitude of states represented in the cohort to a broad geographic and the wide departmental interest in having the skills available demonstrate the significant need – perhaps in regional clusters through an empowered

approach. Horizontally, interest in cybersecurity skills and detectives/investigators and patrol officers. This interest extended into the domain of police lieutenants and sergeants, with interest also shown by law enforcement officers at the captaincy level as well as by non-sworn personnel. A surprising number of police chiefs and assistant chiefs were also represented in the training.

Vertically, there was learner interest from 107 departments (with a few providing several law enforcement officials from their departments) coming from in over 100 different cities across 33 states, from Alabama to Washington state. This supports the wide interest and the need for the services across law enforcement in the United States.

Training has come in the basic series of five courses that cover IT basics, legal issues, network security, information security, and computer forensics. Although all of the initial learners across the three cohorts are participating only for training, we will review the impact of continuing law enforcement education accreditation for the program. Kentucky Law Enforcement Council approved the basic training series or law enforcement continuing education, which will also be available to other departments around the United States under reciprocity agreements. For example, Kentucky accepts reciprocity from law enforcement training on a case-by-basis upon submission of the application for reciprocity credit and information regarding the applicant's work history. Project participants will be advised to contact the training officer for their department as to reciprocity application per state law. For example, Arkansas's Commission of Law Enforcement Standards and Training, Rule 132.0 0.17-001, 1008 Waiver for Equivalent Training, provides "(4) The Commission is authorized to enter into standing reciprocity compacts or agreements with those states which by law regulate and supervise the quality of law enforcement training and which require a minimum number of hours of classroom training in the Basic or Recruit Course equivalent to standards established by the Commission" (Arkansas Commission on Law Enforcement Standards 2021). Another example is the Texas Commission on Law Enforcement, which accepts Out of State Service/Training Credit for Peace Officers (Texas Commission on Law Enforcement 2021). A list of training accreditation sites by state is available through Lenoir Community College, at <https://www.lenoircc.edu/pdf/SA-BLET.pdf> (Accessed 5 December 2021).

These results and the data they produce support the proposition that law enforcement cybersecurity training is desired by state and local law enforcement in the United States and has the potential of increasing distributed cybersecurity across the United States.

4. Conclusion

The wide participation in these projects across multiple dimensions—from law enforcement and public safety to healthcare to national defense—reflects an underlying sentiment for many to protect those in their communities from all threats, to the extent they can.

The diversity of skills, backgrounds, positions, and education in those participating in this new learning demonstrates a broad and deep interest in this new frontier. Further, the participants' commitment to take on additional workloads of cybersecurity education in addition to their full-time jobs demonstrates a commitment to mastering new knowledge to enhance their skills.

Such opportunities must expand to build an adequate cybersecurity regime to protect all of our fellow citizens. These training programs can advance formal credentialing through the development of a structured curriculum that is translatable to a broad array of education contexts, and it will build a distributed security regime to protect and empower more and more people in our United States.

Acknowledgments

We wish to thank the U.S. National Security Agency, the U.S. Department of Defense, the Commonwealth of Kentucky, and the U.S. Department of Homeland Security for their support in making these pilot projects possible in pursuit of the cybersecurity of all Americans. This research was funded by a National Centers of Academic Excellence in Cybersecurity grant H98230-20-1-0347, which is part of the National Security Agency. Continued activities within the online law enforcement training program are conducted with DHS S&T under contract number 70RSAT20CB000021. In addition, we express our appreciation and impactful partnership with the U.S. Office of Economic Adjustment, Department of Defense (DOD) and the Kentucky Commission on Military Affairs

for the Cybersecurity, Certifications, Careers, and Communities (C4) project under the Federal Award Identifier Number HQ00052110003.

References

- Amor, A. M., Xanthopoulou, D., Calvo, N., & Vázquez, J. P. A. (2021). 'Structural empowerment, psychological empowerment, and work engagement: A cross-country study.' *European Management Journal*, 39(6), pp. 779-789.
- Arkansas Commission of Law Enforcement Standards and Training. (2021) *Rule 132.0 0.17-001, 1008 Waiver for Equivalent Training* [Online]. Available at <https://www.law.cornell.edu/regulations/arkaNCAEs/Ark-Admin-Rules-132-00-17>, [Accessed 5 December 2021].
- Ayres, T. (2021) "A Maritime Solution for Cyber Piracy", [online], Wall Street Journal, <https://www.wsj.com/articles/a-maritime-solution-for-cyber-piracy-11620922458>.
- Boeke, S. (2018). 'National cyber crisis management: Different European approaches.' *Governance*, 31(3), 449-464.
- Chow, K.P., Losavio, M., Joshua, J., and Koltay, A. (2018) "The Internet of Things and the Smart City: Legal Challenges with Digital Forensics, Privacy and Security", *Security and Privacy*, 1(3).
- Collier, K. (2021) *Baby died because ransomware attack on hospital, suit says* [Online]. Available at <https://www.nbcnews.com/news/baby-died-due-ransomware-attack-hospital-suit-claims-rcna2465> [Accessed 28 November 2021].
- Cybersecurity and Infrastructure Security Agency. (2020) *Ransomware Guide* [Online]. Available at https://www.cisa.gov/sites/default/files/publications/CISA_MS-ISAC_Ransomware%20Guide_S508C.pdf [Accessed 28 May 2021].
- Cybersecurity and Infrastructure Security Agency. (2021) *Stop Ransomware* [Online]. Available at <https://www.cisa.gov/ransomware> [Accessed 11 Jan. 2022].
- Elmaghraby, A.S. and Losavio, M.M. (2017) Smarter Cities Need Smarter Security. In T. Saadawi and J. Colwell, eds. *Cyber Infrastructure Protection, Vol III*. Carlisle, PA: US Army War College Press. pp 137-156.
- Leong, C. M. L., Pan, S. L., Ractham, P., & Kaewkitipong, L. (2015). 'ICT-enabled community empowerment in crisis response: Social media in Thailand flooding 2011.' *Journal of the Association for Information Systems*, 16(3). <https://doi.org/10.17705/1jais.00390>
- Losavio, M. and Elmaghraby, A. (2021) Artificial Intelligence, Smart Technologies and the Internet of Things: Caution with the Safety, Security and Privacy Eco-system. In Rocha, A. and Isaeva, E. (eds.) *Science and Global Challenges of the 21st Century - Science and Technology Perm Forum 2021*, 18-20 October 2021, Perm, Russia: Springer, Cham. pp 92-102. Available from https://link.springer.com/chapter/10.1007/978-3-030-89477-1_10#citeas.
- Losavio, M.M., Keeling, D., and Elmaghraby, A. (2009) A Distributed Triage Model for Digital Forensic Services to State and Local Law Enforcement. In *2009 Institute of Electrical and Electronic Engineers (IEEE) Fourth International Workshop on Systematic Approaches to Digital Forensic Engineering*, 21 May 2009, Berkeley, CA: IEEE. pp 36-37. Available from <https://ieeexplore.ieee.org/document/5341551>.
- Losavio, M.M., Shutt, J., Keeling, D. (2011) The Information Polity: Social and Legal Frameworks for Critical Cyber Infrastructure Protection. In: T. Saadawi and L. Jordan, eds. *Cyber Infrastructure Protection*. Carlisle, PA: US Army War College Press. pp 129-158.
- Loughnane, J. (2020) *ABA Techreport 2020: Cybersecurity* [Online]. Available at <https://www.lawtechnologytoday.org/2020/10/techreport-2020-cybersecurity/> [Accessed 11 Jan. 2022].
- Paulson, S. L. (1988). An empowerment theory of legal norms. *Ratio Juris*, 1(1), pp. 58-72. <https://doi.org/10.1111/j.1467-9337.1988.tb00004.x>
- Perkins, D. D., & Zimmerman, M. A. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology*, 23(5), 569-579. <https://doi.org/10.1007/BF02506982>
- Poynder, R. (2001) "The Open Source Movement: Does This Software Provide a Viable, User-Friendly Alternative to Proprietary Solutions?", [online], Information Today, <https://www.infotoday.com/it/oct01/poynder.htm>.
- Scott, P. (2020) *Vermont Army National Guard Cyber Response Team to Support UVM Health Network in Response to Ongoing Information Technology System Disruption* [Online]. Available at <https://governor.vermont.gov/press-release/vermont-army-national-guard-cyber-response-team-support-uvm-health-network-response> [Accessed: 11 Jan. 2022].
- Shankar, AJ. (2021) *Ransomware Attackers Take Aim at Law Firms* [Online]. Available at <https://www.forbes.com/sites/forbestechcouncil/2021/03/12/ransomware-attackers-take-aim-at-law-firms/?sh=795f79eba13e> [Accessed 11 Jan. 2022].
- Sun, J. C. (2014). Legal Issues Associated with Big Data in Higher Education: Ethical Considerations and Cautionary Tales. In: Lane J. E. (ed.), *Building a Smarter University: Big data, Innovation, and Ingenuity*. Albany: SUNY Press. Chapter 2.
- Texas Commission on Law Enforcement. (2021) *Out of State Service/Training Credit or Peace Officers... Application* [Online]. Available at https://www.tcole.texas.gov/sites/default/files/FormsAppsPubs/Service%20and%20Training%20Credit_11.30.2021.pdf [Accessed 5 December 2021].
- Wilkinson, A. (1998), 'Empowerment: Theory and practice.' *Personnel Review*, 27(1), pp. 40-56. <https://doi.org.echo.louisville.edu/10.1108/00483489810368549>

On Explainable AI Solutions for Targeting in Cyber Military Operations

Clara Maathuis

Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: Nowadays, it is hard to recall a domain, system, or problem that does not use, embed, or could be tackled through AI. From early stages of its development, its techniques and technologies were successfully implemented by military forces for different purposes in distinct military operations. Since cyberspace represents the last officially recognized operational battlefield, it also offers a direct virtual setting for implementing AI solutions for military operations conducted inside or through it. However, planning and conducting AI-based cyber military operations are actions still in the beginning of development. Thus, both practitioner and academic dedication is required since the impact of their use could have significant consequences which requires that the output of such intelligent solutions is explainable to the engineers developing them and also to their users e.g., military decision makers. Hence, this article starts by discussing the meaning of explainable AI in the context of targeting in military cyber operations, continues by analyzing the challenges of embedding AI solutions (e.g., intelligent cyber weapons) in different targeting phases, and is structuring them in corresponding taxonomies packaged in a design framework. It does that by crossing the targeting process focusing on target development, capability analysis, and target engagement. Moreover, this research argues that especially in such operations carried out in silence and at incredible speed, it is of major importance that the military forces involved are aware of the following. First, the decisions taken by the intelligent systems embedded. Second, are not only aware, but also able to interpret the results obtained from the AI solutions in a proper, effective, and efficient way. From there, this research draws possible technological and human-oriented methods that facilitate the successful implementation of XAI solutions for targeting in military cyber operations.

Keywords: cyber operations, cyber weapons, military operations, targeting, artificial intelligence, explainable AI

1. Introduction

"From this place, and from this day forth, begins a new era in the history of the world, and you can all say that you were present at its birth." (Goethe)

Crossing centuries through the classical OODA (Observe, Orient, Decide, Act) loop, a long series of wars were conducted since the 17th century, spanning the local and world wars, and going into today's and future's wars. Accordingly, the Observe topic moved from telescope (wars in the 17th century) to radio and radar (WWII) and goes to network (future wars). The Orient topic moved from weeks (wars in the 17th century) to hours (WWII) and is going to be continuous (future wars). The Decide topic moved from months (wars in the 17th century) to days (WWII) and will be immediate (future wars). The Act topic transformed to according to season (wars in the 17th century), to weeks (WWII), and goes to minutes (future wars) (Lehto, 2016). These developments were possible due to significant technological advancements that the scientific community and industry professionals proposed and continue to do. Among these developments AI finds its place. No matter if used for planning optimization, target identification, or effects assessment, AI showed impressive results in different prediction, simulation, or exploration problems (Samek & Müller, 2019) tackled in cyberspace. Here, the operations are conducted using cyber weapons/capabilities to achieve military objectives inside and/or outside it (Maathuis, Pieters & Van den Berg, 2018a) having (un)foreseen (in)direct effects on targets and collateral entities (Maathuis, Pieters & Van den Berg, 2018b). Compared to other military operations, cyber military operations take place in silence and at high speed, allow early intelligence gathering and preparations, and imply diverse options for building cyber weapons corresponding to targets' vulnerabilities, nature, aim, and context. In these moments, the decisions made and implications considered, have to be clear and understandable to the stakeholders involved.

Centrally located in the ongoing developments and successes of AI is machine learning, and in particular deep learning. While the improvements, accessibility, and use of these paradigms represent a mindset change and a turnover on data focused applications, also imply understanding and dealing with the decisions made and the results obtained by models: a complex and difficult task. However, recently, a starting point in this direction is the DARPA program (DARPA, 2016) aimed at creating human understandable AI models through a set of design options considering the explainability-performance trade-off. This research line captured the interest of academic researchers and practitioners from different fields (Adadi & Berrada, 2018), but a theory for XAI and a

universally agreed definition is lacking (Samek & Müller, 2019). Moreover, especially in this domain, for targeting is important that decision makers are aware which and how data is used since data could be scarce in this battlefield or could come from several battlefields, understand which decisions the model takes inside and outside cyberspace, and are conscious of the results proposed in order to integrate and deploy it properly with minimal or without risks in the operational field since such risks and their corresponding effects could be experienced both digitally and physically and at large scale due to cyberspace's particularities like interconnectivity and dynamic nature. Given the aspects abovementioned, understanding and applying XAI in cyber military operations is a complex task and has to be tackled considering multiple angles. On this matter, to the best of our knowledge XAI was not defined and tackled in the military cyber context, thus we aim to address this having the following objectives:

- To propose a definition and common understanding for XAI in military cyber operations.
- To address challenges of XAI in military cyber operations through techno-military and socio-ethical lenses focusing on the first phases of the targeting process.
- To raise the level of awareness and responsibility of decision makers on the design and implementation of XAI models for military decision support and broader in the military cyber domain.
- To contribute to the design of strategies, standards, and methods for XAI in the military cyber domain. Further, to stress the need for education in XAI for both current and future decision makers.

To achieve these objectives, multidisciplinary research is conducted based on extensive literature review and analysis. Furthermore, the contributions of this research are twofold. First, to merge and highlight various aspects that should be considered when grasping, developing, and evaluating XAI with a focus on three phases of targeting in military cyber operations. And second, to serve as a design framework with concrete recommendations for developments for decision makers.

The remainder of this article is structured as follows. Section 2 discusses related research concerning theoretical and practical aspects of XAI. Section 3 tackles the targeting process while reflecting on the phases that this research focuses on. Section 4 addresses the necessity of proposing XAI models, reflects on different types of participating stakeholders, analyses characteristics of good explanations, and advances a definition for XAI in this domain. Section 5 considers different types of explanations, explanation methods, and evaluation criteria and mechanisms for explanation methods. Section 6 tackles challenges for XAI for targeting in military cyber operations. Section 7 reflects on the findings of this research and discusses future ideas.

2. Related Research

The general interest in developing AI techniques significantly grew in the last decade, and this is of particular interest in respect to proposing both theoretical and practical approaches for XAI in different domains.

Concerning the meaning and development of XAI, Arrieta et al. (2020) summarize previous efforts and consider that XAI is "widely acknowledged as a crucial feature for the practical deployment of AI models" and position the audience of such models being a key element when understanding it. Accordingly, one of the reflection points is the audience and stakeholders involved in XAI. Preece et al. (2018) attribute the fact that there is no consensus for explainability and interpretability to the fact that different stakeholder communities have to deal with them, and further depict where this perspective overlap and where not. Moreover, Samek, Wiegand & Muller (2017) identify as reasons for needing XAI the following: system verification, learning from the system, and compliance to legislation. Gerlings, Shollo & Constantiou (2020) use socio-technical lenses for analyzing technical and governance aspects for XAI implementation e.g., compliance with regulation and GDPR plus bias and misinterpretation minimization. Hereof, Mohseni, Zarei & Ragan (2021) analyze different explanations and what should be explained by models further advancing a series of methods e.g., explanation satisfaction measured using criteria like user satisfaction and explanation usefulness. Thusly, Vilone & Longo (2020) provide an extensive review on explanation types, explanation methods, evaluation strategies, and future ideas for research like human-in-the-loop approach and interactive interfacing. As initiator of XAI, DARPA (2016) built a successful program finalized in 2018 with different teams working on technical aspects in the design, implementation, and evaluation of XAI models, and a team working on finding, defining, and applying psychology theories of explanation. Furthermore, the vision of the European Commission is discussed by Hamon, Junklewitz & Sanchez (2020) regarding transparency of AI models: i) documenting the AI processing chain using the technical principles of the model plus the data representations used for its design, ii) reliability of AI models that relates to their capacity of avoiding failure or malfunction due to edge cases or malicious intentions; and iii) data

protection in models for preserving security and managing risks through technical and organizational controls. Additionally, humans have the right to obtain an explanation as stressed in Recital 71 of the GDPR (Hamon, Junklewitz & Sanchez, 2020). Moreover, the U.S. DoD adopted a series of ethical principles for AI applied in (non)combat functions for upholding legal, and ethical, and policy: responsible, equitable, traceable, reliable, and governable (U.S. DoD Board, 2019).

As applications, Streich et al. (2020) analyze XAI's potential for tackling technological issues regarding producing sustainable agriculture systems corresponding to the UN sustainable development goals, and Shukla, Fan & Jennions (2020) generate guidelines for building XAI for Integrated Vehicle Health Management clearly understood by human experts using information for health assessment of the subsystems and their aircraft effects. In the medical domain, Holzinger et al. (2017) tackle the possibility of developing methods that reenact the machine decision-making process taken by XAI models in different processes. In the industry sectors, Guo (2020) argues that explainability enables trust i.e., a critical quality for 6G technology since is managing a wide range of mission critical services like autonomous driving, and Lai et al. (2020) implement a XAI model with the underlying physics of high-energy particle collisions using information encoded in the energy-momentum four-vectors of the final state particles using GANs.

In the military domain, Bistrion & Piotrowski (2021) discuss applications embedding AI techniques and their impact on security sensing and show that military applications are among the ones responsible for AI development. Furthermore, Preece et al. (2019) map ISR requirements in multi-domain operations with the need of building XAI models to produce robust human-machine decision making with examples from urban terrain analysis and enhanced asset interoperability, and Hepenstal & McNeish (2020) argue that when designing military and security XAI solutions, careful consideration on the context and nature of the problem being modelled and the humans involved should be taken. In particular, Maathuis, Pieters & Van den Berg (2020) propose a multi-layered fuzzy XAI model for effects assessment and decision support for targeting in military cyber operations, and Keneni, et al. (2019) present a XAI model that depicts UAVs decisions' logic in a predefined mission.

These resources capture important aspects that should be considered when defining and proposing AI solutions in the military cyber domain but are not tailored to it and this is necessary for building trustable, explainable, and accountable military intelligent systems designed, implemented, and used in the execution of cyber military operations whose effects may by their nature cross geographical or digital boundaries. It is then the aim of this article to tackle this knowledge gap and propose a definition and model for XAI in the military cyber domain, reflect on its corresponding challenges, and further discuss possible solutions.

3. Military Targeting

To fight adversaries and achieve goals, cyber military operations are conducted by influencing their target(s) (embedding ICT elements or direct cyber targets) in several ways e.g., disrupt communication processes or alter the behavior of an audience. In this process, particularities and challenges like the dual-use nature, connectivity, dynamism, uncertainty, and attribution issue existing in cyberspace should be considered when building AI-based solutions and represent a direct reason for developing XAI-based solutions.

At the core of conducting cyber military operations is the military targeting process and corresponding Rules of Engagement (RoE) that establish the circumstances and limitations for engagement in each operation. The targeting process is defined by (NATO, 2016; U.S. Army, 2018) as selecting and prioritizing targets and matching appropriate response to them while considering operational requirements and capabilities. This process contains six phases as depicted in Figure 1 and below summarized (NATO, 2016; U.S. Army, 2018; Boothby & Schmitt, 2012):

- Phase I (Commander's intent, objectives, and guidance): define clear objectives and under which circumstances and actions these objectives should be achieved, political and strategic guidance is provided, further operational tasks are created, and targets are nominated.
- Phase II (Target development): eligible targets are analysed, vetted, validated, and prioritized resulting in a prioritised target list. Herein are also included collateral damage estimation and intelligence gain/loss assessment which implies that commanders might not engage a target in order to negatively influence the process but would benefit from the results during target selection.

- Phase III (Capability analysis): the targets are analysed and matched with capabilities to produce the desired effects while minimizing collateral damage i.e., proportionality assessment and further CoA (Course of Action) development.
- Phase IV (Commander's decision, force planning, and assignment): based on the results obtained, further assignments to specific forces are done for planning and execution considering any relevant constraints and restraints.
- Phase V (Mission planning and force execution): the mission is further planned and executed at tactical level while a final target PID (Positive Identification) is conducted with other information checks and collateral damage avoidance or minimization.
- Phase VI (Assessment): the effects produced are evaluated next to the achievement of the objectives defined, further contributing to broader assessments, input for other operations, and lessons learned.



Figure 1: Military targeting process

To go through these phases, vast amounts of procedures, methods, and models are applied merging diverse sources and types of data i.e., system, process, and human. No matter if data are further processed and used to build simulation, prediction, or combined intelligent models, models' decisions, output, and possible impact should be properly understood and justified by military decision makers. It is in this context that XAI is further defined and analysed considering its challenges and possible solutions.

4. Defining Explainable AI

Although it sounds as one of the major topics in the field of AI in the last decade, XAI finds its roots five decades before. Since then, the importance of explainability and interpretability of intelligent systems it actually increased (Hansen & Rieger, 2019) and returned as a focus point due to the increased developments in machine learning, in deep learning which although proved its performance in a diverse plethora of tasks, is still intrinsically unable to explain its decisions in a human understandable way. In other words, as Preece, et al. (2018) stress: "it is not a new problem, nor was it ever considered a solved problem". But why XAI in military cyber operations? More concretely, what are the reasons and objectives of XAI? The following reasons are considered:

First, military cyber operations that use intelligent systems could have (in)direct, (un)intended, and (un)expected effects not only on their targets, but also on collateral assets. Especially in regards with the unintended effects, properly *understanding* where a target is localized physically and digitally, and which of its vulnerabilities is more prone to exploit in an early operational state, what is the intelligence gain/loss, and how is selected and prepared for engagement (Phase II-III) could facilitate the successful deployment of an intelligent cyber weapon (XAI based) in Phase V.

Second, coupled with understanding why and how an AI model is implemented and used, is important to *comprehend the decisions* made by a model and further *communicate* them to the systems, processes, and people involved. In particular, if a proper match to a cyber weapon is done in Phase III, its choices should be accordingly argued and communicated for execution plans and engagement from Phases IV-V.

Third, since targeting is an ongoing process carried out by multiple teams, is important that the functional mechanisms and decisions made by the AI models used (e.g., for target localization, cyber weapon selection and, effects estimation) are *justifiable* and transparent to facilitate and strengthening *control, trust, and accountability* (Adadi & Berrada, 2018; Burkart & Huber, 2021) along the entities involved. Moreover, this allows direct correspondence and mapping to legal checks, strategic goals, and socio-ethical values.

And the list of reasons could continue depending on the aim, activities, plus the granularity considered and appraise the uncertainty, dynamism, and strong interconnectivity aspects that characterize cyberspace. However, as these reasons reflect, XAI is first about understanding while maintaining a high level of performance (Gunning & Aha, 2019), secondly about sustaining proper communication in a human friendly way, and thirdly about justification, trust, and accountability in the systems used. As Russell & Norvig (2021) consider: “explanations are not decisions, they are stories about decisions”. Thusly, we take the stance of (Molnar, 2020) that humans “desire to find meaning in the world”, and after assessing that the word ‘explainable’ is not defined in Oxford (Oxford Dictionary, 2021) and Cambridge (Cambridge Dictionary, 2021) dictionaries, we see that the closest term is ‘explanation’ meaning a reason, a fact, or an excuse for something. Besides, for defining XAI in military cyber operations, we align with the definition perspectives of (DARPA, 2016; Arrieta et al., 2020; Doran et al., 2017) and consider XAI as:

XAI in the cyber military domain = a sub-field of AI that deals with the design, development, and use of methods, techniques, and technologies that provide reasons and facts for the functional mechanism, decisions, and results made by AI systems embedded in different cyber military systems and processes.

This definition could be reduced to a series of (cyber) agents that interact with each other and make use of a series of tools to produce actions. In Figure 2 we illustrate this definition while explaining its components:

Agents: entities participating in the design, development, and use of XAI in cyber military operations. Seeing their context, nature, role, and involvement in this process, they can be classified as (DARPA, 2016; Meske et al., 2021; Arrieta et al., 2020; Preece et al., 2018; Hepenstal & McNeish, 2020; Hamon et al., 2020):

- *Stakeholders*: entities involved either i) in the process of design, standardization, and certification of the models i.e., military-legal and AI regulators which should also integrate ethical aspects, or ii) in the process of theorizing, designing, developing, debugging, validating and upgrading the models i.e. AI and military cyber engineers, or iii) in the process of design, use, and assuring compliance of AI systems in and outside the organization i.e. AI managers and cyber military decision makers.
- *Audience*: entities directly involved in the processes done by stakeholders i.e., are participatory audience or end-users i.e., they represent the other audience.

Tools: methods, techniques, and technologies used by agents to design, develop, and use XAI in cyber military operations (DARPA, 2016):

- *XAI model* is the developed AI model by stakeholders and (perhaps) participatory audience.
- *XAI interface* is the interface that generates explanations i.e., actions to the audience. Accordingly, it answers to questions like why? how? and when? e.g. Why this or why not something else? How is it done or how to tackle or correct errors? When is it successful, when not, or when will fail?

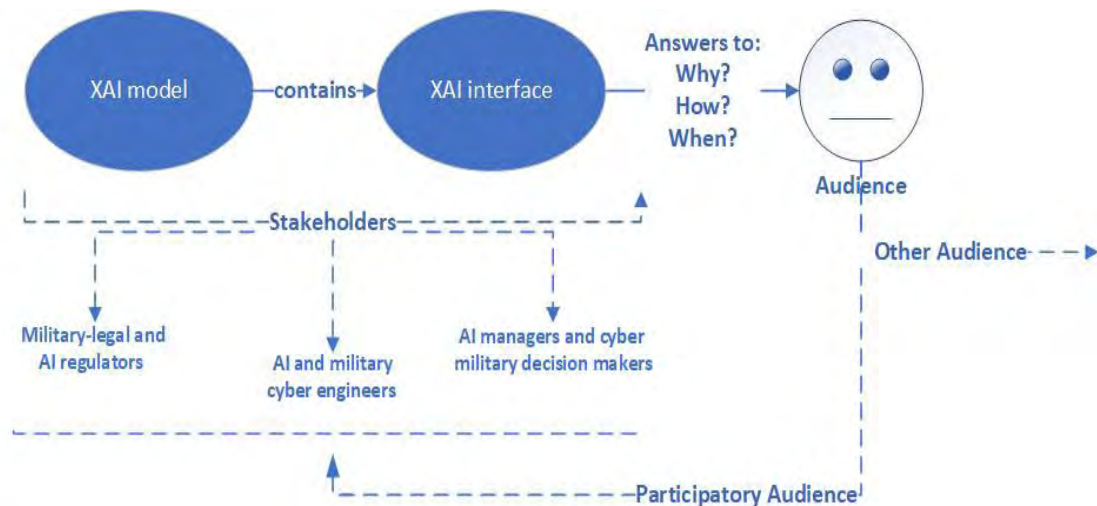


Figure 2: XAI in cyber military targeting, based on (DARPA, 2016; Meske et al., 2021)

To assess proportionality for in-house implemented military cyber operation, all stakeholders cooperate for building a XAI cyber weapon for assessing the expected collateral damage and military advantage in a human understandable way through a direct or developed XAI interface (depending on the technique) and provide a decision, meaning is this engagement proportional or disproportional, understandable, and transparent to the audience i.e., military decision makers. As seen above, a plausible question would then be: Why is proportional/disproportional? In case that the decision to engage a target is negative, then the intelligence gain/loss assessment benefits from clear explanations regarding target's location, vulnerabilities, and connections i.e., information beneficial during the ongoing or future operation(s). Then, a relevant question would be: What information could be further used?

5. Explainable AI Methods and their Evaluation

When analysing explanations regarding the role of the involved stakeholders and audience, one could consider the intention of the explanation method that refers to which question is answered, and the intention of the one(s) that use the explanation meaning how and for what should be this explanation used (Samek & Müller, 2019). Furthermore, Islam, Eberle & Ghafoor (2020) consider that explanations should be expressive, translucent, portable, accurate, have fidelity, consistent, stable, comprehensible, and contain a degree of importance. Hence, the following criteria are considered to classify the existing methods for XAI (Vilone & Longo, 2020; Samek & Müller, 2019; Arrieta et al., 2020; Kolbasin, 2018):

- *XAI model* is the developed AI model by stakeholders and (perhaps) participatory audience.
- *Problem type*: reflects the problem studied e.g., classification or regression.
- *Scope*: the objective of an explanation: i) local which implies that each inference of a model is explained e.g., a specific rule for proportionality assessment, while ii) global implies that the complete inferential process of a model is transparent and understandable as a whole e.g., general mechanism for proportionality assessment.
- *Stage*: the moment when explanations are generated: i) *ante-hoc* methods imply explainability from the beginning and during the training phase using concrete examples e.g. vulnerabilities that a target has or concrete rules that determines if engaging a target is disproportional or not, while ii) *post-hoc* methods imply mimicking model's behaviour using an external explainer during the testing phase e.g. one that would explain target's engagement with a cyber weapon or proportionality assessment rationality. These methods could be either a) *model agnostic* applied to any type of model, the model is treated as a black box, and the explanations are generated without the inspection of internal parameters, and b) *model specific* limited to specific types of models embedding specific model logic.
- *Input data*: depending on aspects like availability, volume, and data type (e.g., numerical, categorical), the explanations are considered.
- *Output format*: depending on the format of desired results (numerical, rules, textual, visual, or mixed), the explanations are considered.
- *Moment of providing explanations*: i) *before building the model* using methods like exploratory data analysis using different visualization methods, understanding distributions, or features analysis e.g. multiple entry points could be analysed for a series of targets to find out their centre of gravity and build

a corresponding cyber weapon in Phases III-IV, ii) *during building the model* directly for models like decision trees or linear ones e.g. the assessment of proportionality could be explicitly provided with its corresponding rules and results in Phases IV-V, and iii) *after building the model* that explain the model, its outcome, and internal processes for models based on neural networks and SVM e.g. in case of network analysis for possible target inspection done in Phase II.

However, one needs to make sure that the actions of AI models do not only meet their objectives, but actually human objectives (Russell, 2019) to consider them beneficial and trustable. These facts should be verified in both cases where AI models have a supportive or a decisive role (Samek & Müller, 2019). Accordingly, the quality of explanations depends on how the stakeholder/audience perceives them depending on their background, goals, expectations, context etc. (Atakishiyev et al., 2020). Hence, based on (Swartout & Moore, 1993; Walsh et al., 2021), the following criteria are considered to evaluate explanations of AI models:

- *Measure of suitability, sufficiency, and fidelity*: the explanations provided should be suitable, sufficient, and representative for system's actions.
- *Measure of effectiveness*: the explanations provided should be clear, useful, and understandable for the stakeholders and audience of the system.
- *Measure of performance*: the explanations provided should improve the current perceptions, beliefs, and perhaps competences that stakeholders and audience have on the system while making sure that this would not overhead or slow down the system.

6. Challenges of Explainable AI

As any kind of disruptive technology, XAI, presents a series of challenges in the military cyber domain. For structuring and characterizing them, socio-technical lenses are considered to capture both technical and socio-ethical elements. Thereupon, the following challenges are considered:

- Insufficient (training) data (Svenmarck et al., 2018): insufficient or not representative (training) data has a direct impact on the results provided by a model which could have significant negative consequences e.g., wrong target considered for engagement.
- Data ownership, protection, sharing, availability, and quality (Maxwell, 2020; Stahl, 2021): not only that data should be in the hand of its owners which have the duty to protect or share when and to who is necessary but should be available in different stages and maintain its quality during the processing and modelling phases.
- Loss of debuggability and transparency in development and testing (Google, 2020).
- Performance, maintenance, and robustness costs (Core et al., 2006; Gunning & Aha, 2019; Walsh et al., 2021).
- Lack of control (Google, 2020) for both stakeholders and audience.
- AI security attacks (Svenmarck et al., 2018; Morgan et al., 2020): since AI models are software-based systems, they are vulnerable to well determined adversaries that could conduct e.g., poisoning, adversarial, or differential attacks. For instance, by altering training data used for target engagement, another object could be engaged and/or the produced collateral damage could be higher than expected in Phases III-V.
- Real-live settings mirroring environment (Stahl, 2021): used for evaluating and simulating the model in realistic conditions that capture actual aspects using different simulation settings like test beds and digital twin solutions.
- Alignment with strategic and operational needs (Walsh et al., 2021): since conducting cyber military operations is a set of processes carried out by members of teams with different expertise, an alignment between their objectives and approach could be sometimes challenging.
- Alignment with commercial developments and standards (Hoadley & Lucas, 2018; Walsh et al., 2021) e.g., industry related cyber systems and solutions.
- Integration with existing systems and emerging capabilities (Walsh et al., 2021) from other military domains.
- Need for developing corresponding international political, legal, and military initiatives, standards, strategies, and methods, and further compliance with them (Deeks, 2019; Hoadley & Lucas, 2018; Samek & Müller, 2019; Morgan et al., 2020).
- Invest in education, building human capacity, and R&D (Alonso, 2020; Morgan et al., 2020).
- Fostering and building a digital ecosystem for AI (Morgan et al., 2020).

- Creation of International competition, countries asymmetries, and power imbalance (Hoadley & Lucas, 2018).
- Negative impact on human values, health, environment etc.
- Uncertainty due to the dynamism of cyberspace and the unpredictability character of conflict (Maathuis, Pieters & Van den Berg, 2016).

7. Conclusion

The mantra of AI continues to change from the more developing intelligent systems the better (Russell, 2019) to the more useful, explainable, and responsible developing intelligent systems the better. This is a long path for the scientific and industry communities from all societal domains. Of special interest, considering the potential of AI to increase the likelihood of war, to escalate ongoing conflicts, and to proliferate to malicious actors (Morgan et al., 2020) plus the new operational cyber tools that have the potential of digitally engaging their adversaries in silence and at incredible speed, it is the duty of decision makers involved in developing and/or conducting such operations, to make sure that AI is properly understood, rightly used, and that the positive aspects are promoted while taking into account, mitigating, or avoiding the negative ones (Russell, S., & Norvig, 2021). These aspects foster trust in AI systems by allowing their users to understand their behavior, limitations, and post-mortem assessment (Dignum, 2019). Addressing these aspects in the military cyber domain, this research aims at defining and analysing essential elements and methods relevant for the design and application of XAI models when targeting in military cyber operations. Hence, this research contributes to the existing body of knowledge in the AI, military, and cyber security domains, and calls for further research in this direction while also representing a basis for further design of strategies, policies, methods, and techniques applicable in the military cyber domain.

This research continues by analysing the challenges, opportunities, and effects of developing XAI models in the military cyber domain in a multidisciplinary setting with concrete cases and modelling solutions. Additionally, this research stresses the need for stakeholders' education and awareness with dedicated programs and courses, and the need of developing modelling and simulation, gaming, test beds, and digital twin solutions that facilitate and strengthen the responsibility, transparency, and fairness of i) the stakeholders involved in developing XAI models in the military cyber domain, and ii) the XAI models themselves.

References

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*.
- Alonso, J. M. (2020). Teaching Explainable Artificial Intelligence to High School Students. *International Journal of Computational Intelligence Systems*, 13(1), pp. 974-987.
- Arrieta, A. B. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp. 82-115.
- Atakishiyev, S. et al. (2020). A multi-component framework for the analysis and design of explainable artificial intelligence. *arXiv preprint arXiv:2005.01908*.
- Boothby, W., H. and Schmitt, M., N. (2012). *The law of targeting*. Oxford University Press.
- Bistrón, M. and Piotrowski, Z. (2021). Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. *Electronics*, 10(7), 871.
- Board, D. I. (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the *Department of Defense: Supporting Document*.
- Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7), pp. 1829-1850.
- Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, pp. 245-317.
- Cambridge Dictionary (2021). <https://dictionary.cambridge.org/dictionary/english/explanation>
- Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S. and Rosenberg, M. (2006). Building explainable artificial intelligence systems. In *AAAI* (pp. 1766-1773).
- DARPA. (2016). Explainable Artificial Intelligence. <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.
- Doran, D., Schulz, S. and Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.
- Gerlings, J., Shollo, A. and Constantiou, I. (2020). Reviewing the Need for Explainable Artificial Intelligence (xAI). *arXiv preprint arXiv:2012.01007*.
- Google (2020). AI Explainability Whitepaper. *Google*. European Conference of the Prognostics and Health Management Society.
- Gunning, D. and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), pp. 44-58.

- Guo, W. (2020). Explainable artificial intelligence for 6G: Improving trust between human and machine. *IEEE Communications Magazine*, 58(6), pp. 39-45.
- Hamon, R., Junklewitz, H. and Sanchez, I. (2020). Robustness and explainability of artificial intelligence. *Publications Office of the European Union*.
- Hansen, L. K. and Rieger, L. (2019). Interpretability in intelligent systems—a new concept?. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, pp. 41-49
- Hepenstal, S. and McNeish, D. (2020). Explainable Artificial Intelligence: What Do You Need to Know?. In *International Conference on Human-Computer Interaction*, Springer, pp. 266-275.
- Hoadley, D. S. and Lucas, N. J. (2018). Artificial intelligence and national security.
- Holzinger, A., Biemann, C., Pattichis, C. S. and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923*.
- Islam, S. R., Eberle, W. and Ghafoor, S. K. (2020). Towards quantification of explainability in explainable artificial intelligence methods. In *The Thirty-Third International Flairs Conference*.
- Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaiantz, J. D. and Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access*, 7, pp. 17001-17016.
- Kolbasin, V. (2018). *How to explain predictions of your network?* Globallogic.
- Lai, Y. S., Neill, D., Płoskoń, M. and Ringer, F. (2020). Explainable machine learning of the underlying physics of high-energy particle collisions. *arXiv preprint arXiv:2012.06582*.
- Lehto, M. (2016). Theoretical examination of the Cyber Warfare environment. *Proceedings of the International Conference on Cyber Warfare and Security*, pp. 223-230.
- Maathuis, C., Pieters, W. and Van Den Berg, J. (2016). Cyber weapons: a profiling framework. In *International Conference on Cyber Conflict 2016 (CyCon US)*, IEEE, pp. 1-8.
- Maathuis, C., Pieters, W. & Van Den Berg, J. (2018a). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security (ECCWS)*, pp. 278-88.
- Maathuis, C., Pieters, W. and Van Den Berg, J. (2018b). Assessment methodology for collateral damage and military (dis)advantage in Cyber Operations. In *International Conference on Military Communications (MILCOM)*, IEEE, pp. 1-6.
- Maathuis, C., Pieters, W. and Van den Berg, J. (2020). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*. <https://doi.org/10.1016/j.dt.2020.04.007>.
- Maxwell, P. (2020). Artificial Intelligence is the future of warfare (just not in the way you think), <https://mwi.usma.edu/artificial-intelligence-future-warfare-just-not-way-think/>
- Meske, C., Bunde, E., Schneider, J. and Gersch, M. (2021). Explainable intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, pp. 1-11.
- Mohseni, S., Zarei, N. and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11, pp. 1-45.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
- Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K. and Grossman, D. (2020). *Military applications of artificial intelligence: ethical concerns in an uncertain world*. RAND.
- NATO (2016). *NATO Standard AJP-3.9 Allied Joint Doctrine for Joint Targeting*. NATO Standardization Office.
- Oxford Dictionary (2021). <https://www.oxfordlearnersdictionaries.com/definition/english/explanation>
- Preece, A., Harborne, D., Braines, D., Tomsett, R. and Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.
- Preece, A., Braines, D., Cerutti, F. and Pham, T. (2019). Explainable AI for intelligence augmentation in multi-domain operations. *arXiv preprint arXiv:1910.07563*.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.
- Russell, S. and Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition 4th. *Foundations*, 19, 23.
- Samek, W., Wiegand, T. and Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
- Samek, W. and Müller, K. R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, pp. 5-22.
- Shukla, B., Fan, I. S. and Jennions, I. (2020). Opportunities for Explainable Artificial Intelligence in Aerospace Predictive Maintenance. In *PHM Society European Conference*, 5(1), pp. 11-11.
- Stahl, B. C. (2021). *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, Springer Nature, pp. 124.
- Streich, J. et al. (2020). Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Current opinion in biotechnology*, 61, pp. 217-225.
- Svenmarck, P., Luotsinen, L., Nilsson, M. and Schubert, J. (2018). Possibilities and challenges for artificial intelligence in military applications. In *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting*, pp. 1-16.
- Swartout, W. R. and Moore, J. D. (1993). Explanation in second generation expert systems. In *Second generation expert systems*, Springer, pp. 543-585.
- United States Army Joint Publications 3-12. (2018). *Cyberspace Operations*.
- Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Walsh, M. et al. (2021). *Exploring the Feasibility and Utility of Machine Learning-Assisted Command and Control: Volume 2, Supporting Technical Analysis*. Rand.

Analyzing the Performance of Block-Splitting in LLVM Fingerprinting

William Mahoney¹, Philip Sigillito¹, Jeff Smolinski¹, J. Todd McDonald² and George Grispos¹

¹University of Nebraska at Omaha, Nebraska, USA

²University of South Alabama, Mobile, USA

wmahoney@unomaha.edu

psigillito@unomaha.edu

jsmolinski@unomaha.edu

jtmcdonald@southalabama.edu

ggrispos@unomaha.edu

Abstract: This paper expands and builds upon previous work reported at the 2021 ICCWS concerning Executable Steganography and software intellectual property protection via fingerprinting. Software fingerprinting hides some type of unique identification into the binary program artifact so that a proof of ownership can be established if the artifact turns up elsewhere. In our previous work, it was noted that “fingerprints are a special case of watermarks, with the difference being that each fingerprint is unique to each copy of a program”. This prior work emphasized making the fingerprint independent of the machine architecture; that is, performing the operations on an intermediate representation (IR). LLVM was used as the target IR, which is a compiler “middleware” language that is then converted into machine code in a later step. Both a static fingerprinting method, where the serial number or data is embedded and visible by inspection, and a dynamic method, where the code must be executed, were explored. The dynamic method only incurs an overhead if the proof code is executed and has very minimal impact if the proof code is not executed. However, the static fingerprint was accomplished by shuffling the order of basic blocks in the software in a manner that represents the serial number data, and this would have an impact on both the execution speed and the program size. This paper reports on subsequent research to improve the quantity of data which could be encoded by rearranging the blocks in a program and increasing the number of blocks by splitting them into smaller fragments, thus allowing for more potential orderings and therefore more data. Contributions in the current work are twofold. First, the experimental infrastructure has been refined so that the fingerprinting actions take place within the compiler itself as opposed to an external LLVM parser. Second, code has been introduced to limit the upper bound on the size of a block and to split blocks which are larger than this upper bound. We evaluate the resultant overhead and performance of the block splitting method and report negligible increases based on the block-splitting technique.

Keywords: Security and privacy, application fingerprinting, steganography

1. Introduction

Executable programs can be fingerprinted, making each copy of the software unique. This in turn provides for a certain amount of intellectual property protection, as a copy of the software appearing on eBay – for example – can be traced back to who the software was originally issued to. In fingerprinting, a unique serial number is embedded in some manner into the program. Thus, fingerprinting is a specialized use case for a digital watermark; a watermark can be used to prove authorship while a fingerprint can be used to prove ownership. A fingerprint F is embedded into a program P . There are two alternatives for this process: static fingerprints are detectable by an examination of P ; it is not necessary to run the program to visualize the fingerprint. Contrasting with this is dynamic fingerprinting, where F is caused to appear only when executing a section of P with, for example, specific trigger data.

In a previous paper presented at ICCWS 2021 (Mahoney 2021), fingerprinting methods were described in detail. The paper described work to provide both static and dynamic fingerprinting methods. While the dynamic fingerprinting technique is not reconsidered in this paper, it was triggered by forcing a specific function to execute with a hidden flag set to a certain value. In this prior work the static method uses the serial number as a seed to a pseudo random number generator, which in turn dictates the order of the basic blocks within the program. Basic blocks are sections of executable code which are treated as a unit. Since the ordering of these units is directly impacted by the unique data used to start the random sequence, by examination of the binary artifact the ownership can be proven. All copies of a program will execute with the same expected outcome, but the ordering of the blocks of instructions within each copy is unique.

One aim of the previous work in fingerprinting was to add an additional aim to the set of goals normally encountered in this type of work. These original fingerprinting goals (Colberg and Thomborson 1999) are: 1) the fingerprint should be difficult to remove, 2) the fingerprint should be difficult to detect by the software owner, 3) the fingerprint should have a good data rate and a low false-positive rate, 4) the fingerprint does not severely

impact the performance of the program, and 5) the fingerprint is unique to each instance. The authors add an additional goal of making the fingerprint in such a way that it is language and machine independent. The LLVM intermediate representation in the compilation process is used as the point to interject the fingerprint.

Considering these goals, the ordering of the blocks cannot easily be altered by the program owner without significant investment in time and effort; the fingerprint is difficult to remove. Since the ordering of the blocks is dictated by a one-way function, going from this ordering back to the original serial number is difficult or impossible; the original serial number is not easy to detect. For goal number three, a simple way to think about the data rate is to consider the number of different messages which the program can surreptitiously contain. Of course, the number of possible different messages one can embed into a program by block shuffling depends on the number of blocks. For example, five blocks can carry $5!=120$ different messages, while six blocks can carry $6!=720$ messages. Large programs will be capable of a large quantity of hidden data (goal number three) and because of the growth in this number each program instance can be unique (goal five). This leaves the exploration of the performance impact.

As reported at the previous conference, the authors efforts at fingerprinting in LLVM were successful, but there was a slight impact on the performance of the program. This impact is contrary to goal number four, above. While not significant (4% was measured in the worst case) it begs a question. If one wishes to increase the number of blocks, to carry more messages, what is the resulting impact on the program performance, in terms of space and time? This paper addresses this block size/performance impact question.

The remainder of the paper is organized as follows: a brief description of the LLVM infrastructure is provide in the following section. Section three details other related work in this area, and section four describes previous work in this area by the authors of this paper. Section five contains the results obtained after the previous ICCWS paper. Conclusions and future work are in section six.

2. Background on LLVM

The original intent of the authors is to make the fingerprint in such a way that it is not machine specific. To do this we utilized the LLVM infrastructure. LLVM is the intermediate representation (IR) for several languages, most notably C and C++ but also including Swift, Rust, ActionScript, D, Fortran, Ruby, and many others (LLVM 2021). The LLVM project was started at the University of Illinois at Urbana–Champaign and originally stood for Low Level Virtual Machine (Adve and Lattner 2007). But this is now dropped and LLVM is considered *just* an acronym.

Languages supporting LLVM compile the program source code into the IR. Various tools exist for reading this intermediate code, optimizing it in some way, analyzing it from a machine code perspective, generating files which can be used to graph the control flow or dependence information, generating code analysis information, and other uses (LLVM 2021). Further, the IR can be translated from a bitcode format which can be interpreted, into a human-readable format, and from a human-readable format back to bitcode. The various compiled languages can generate the human-readable format and stop, making for easy inspection and learning about how the intermediate representation works.

Eventually, if needed, the IR is translated into machine code for a particular CPU, and the resulting executable program runs on that machine. This gives language and machine independence for fingerprinting, with the key being that one can compile the program to LLVM and then stop. The IR is then used to embed the data for the fingerprint, and the compilation process finishes by creating the executable program on the target machine. These CPUs supported by LLVM include ARM, x86-32, x86-64, IBM z/Architecture, and many others. LLVM IR can also be used to produce WebAssembly code, which can be interpreted by Microsoft Edge and Google Chrome web browsers. In the case the authors past work, LLVM code was successfully transferred compiled and fingerprinted back and forth between ARM7 and x86-64 CPUs, compiling to LLVM on one machine, fingerprinting, and then finishing on another machine.

3. Related Work

As mentioned above, fingerprinting is frequently either static or dynamic. Both these methods have been extensively studied by various researchers including Colberg (1999, 2002), Nagra (2002), and Palsberg (2000). Reordering of the basic blocks – like that used by the authors of this paper for static fingerprinting – was reported

on as far back as 1996 by Davidson and Myhrvold (Davidson and Myhrvold 1996). Other static fingerprinting methods include data hiding in the interference graphs of register usage (Qu and Potkonjak 1998) and embedding the information in a program's control flow structure (Venkatesan Vazirani and Sinha 2001). Stern has also used modifying the frequencies of the instructions to carry a message (Stern 1999).

In static watermarking, one interesting attempt was to hide executable programs inside of other executable programs by making the operands of certain instructions valid decodable and executable instructions, although the approach is quite limited (Mahoney et al 2018) (Mullins 2020).

In the realm of dynamic watermarking, an approach by Cousot assigns different values to variables as the program executes, and the combination of variable and their value represents the fingerprint (Cousot and Cousot 2004). The branching behavior of programs, the locations that they jump to, has been used to embed the fingerprint so that it is discovered via dynamic analysis (Colberg, Carter et al 2004). Wang, et al. (2018) encode the watermark by causing a program exception and monitoring the exception handling. Academics have also concentrated work in dynamic watermarking by making a software fingerprint distinct in terms of data structures (Kamela and Albluwib 2009) or using graph theory (Bento et al 2019) (Hamilton and Danicic 2020) (Colberg, Huntworth et al 2004).

A new method for dynamic fingerprinting is proposed by Alrehily and Thayananthan (2017) who use Return Oriented Programming (ROP). A consequence of using ROP is that the watermark is dynamic, and by the very nature of the ROP method the watermark is immune to anti-malware efforts such as address space randomization. Using opaque predicates, conditions that are known but must still be evaluated at runtime, Arboit (2002) has watermarked Java programs; this method is further studied by Myles (2006). A novel dynamic method is reported on by Ma et al (2019), using the Collatz conjecture. This conjecture is an unsolved mathematical problem which is used to change the control flow of the program in a detectable manner. Lastly, Preda and Pasqua (2017) explore the semantics of software watermarking as opposed to the methods.

The previous work by the authors of this paper appears to be one of few using the LLVM infrastructure; one other project with similar aims is the LLVM Obfuscator project (Junod et al 2015) which suggests similar benefits in terms of language and CPU independence, but with the focus on making the reverse engineering process more difficult. Their system includes obfuscation techniques such as substituting different instructions with the same semantics, removing cues that are present in the program control flow, and inserting erroneous (but never executed) instruction paths. There are no mentions of the performance impact, however.

4. Expanding Static Fingerprinting

In this follow-up research work the focus is strictly on the prior work with static – as opposed to dynamic – fingerprinting. Since the initial report the authors have modified the method used to introduce the fingerprint, so a review of the previous method and the new method is in order. In both cases the method utilized is based on previous work by Davidson and Myhrvold (1996). The fingerprint is created by modifying the ordering of the basic blocks within a program according to a set rule based upon the serial number of the software artifact. The serial number seeds a pseudo-random number generator which is used to dictate the order of the blocks. This method does not create a fingerprint that is necessarily reversible, but it does uniquely create a fingerprint based up on the executable's unique serial number. And the method does not impact the semantics of the program other than a potential degradation in performance, which is the point of this paper.

In the original implementation, a custom LLVM parser was created that could take a valid LLVM file and parse it into an in-memory parse tree. Programs that were to be fingerprinted were first compiled to LLVM, at which point the process is interrupted for the fingerprinting. The LLVM parser would then create a parse tree in memory and search the parse tree for functions. The types of the nodes within the data structure mimic the names in the grammar, so locating the body of functions is straightforward. The blocks within each function of the program can be determined, put into a collection, and then subjected to a Fisher–Yates shuffle (Fischer 1938) (Knuth 1969), and placed back into the IR tree representation. Once rearranged, a new LLVM intermediate file is created, and the compilation process is allowed to continue using the new IR file.

This initial method was working but turned out – over time – to be fraught with issues, making it less than ideal as a long-term solution. First, the grammar of LLVM seems to continually change from version to version, so

creating a viable grammar for the parsing of the LLVM proves to be like shooting at a moving target. Also problematic, LLVM uses a format called Static Single Assignment (SSA) which assigns unique names to all temporary locations used by the function and enforces the rule that no variable can be used before it is defined. Rearranging the order of the blocks keeps the same semantics but breaks the SSA rule. It is necessary to not only rearrange the blocks but traverse through the function, renaming all variables in a new ordering. But of course, different versions of LLVM have different rules as to what needs or does not need a variable number; every new version of LLVM would break the existing infrastructure.

For these reasons the approach was changed to modifying the actual compiler used – named Clang – to compile C and C++ programs into LLVM. The main advantage to this approach is that it almost singlehandedly eliminates the two problem areas in the prototype. First, if the language grammar for the human readable LLVM changes, one can use the parser integrated into the actual compiler which will of course match the latest changes in the grammar. Secondly, after the blocks are rearranged according to the fingerprint, there are already existing functions within the Clang code to renumber all variables in a function, making what used to be quite complex into a trivial function call. A command line parameter was added to Clang to turn on and off the additional various features to aid with testing.

Once this infrastructure was in place and working, the authors went back to refocus on the research question outlined previously. If one wishes to increase the data carrying capacity of a program, one way is to make blocks smaller. This results in a function having more blocks, therefore more potential permutations of the blocks, and thus a larger message carrying capacity. But this is at the potential cost in performance. Options were added to the modified Clang compiler to shuffle (or not shuffle) the blocks, and to “set the maximum size of a block to X”. This latter rule causes the software to split blocks according to this maximum. A block with, for example, ten LLVM instructions but with a maximum block size set to six, will split the block into two blocks, one of six and one containing the remaining four LLVM instructions. A jump instruction added to the end of the first will transfer execution to the beginning of the second, regardless of what order the blocks might be in when converted to assembly.

5. Case Study

The original aim of the modifications was to split blocks into smaller fragments, to increase the data capacity of the program. The additional number of blocks directly impacts the data capacity by increasing the number of possible orderings of the blocks when they are shuffled. The larger number of blocks, when shuffled, should have a large impact on the performance of the program, both in terms of speed as well as size. To analyze these impacts requires two tools: a standard set of benchmarks and code to analyze the resulting executable programs made from these benchmarks.

5.1 Methodology

For measurements we turned to a subset of the “SPECspeed” benchmarks in the SPEC CPU 2017 benchmark package (SPEC 2017). The subset was selected according to two factors: first, the compiler modifications referred to above were in the Clang compiler and thus appropriate to C or C++ but not Fortran. Second, we want to have a good representation of the shuffling, and thus a program consisting primarily of (or almost exclusively of) external library functions were ignored. These restrictions leave eight benchmarks which seemed applicable for measurements, as well as an overall average measurement among this group.

It was also necessary in the prior study to create a program to analyze x86 instructions. This analysis program was written in Python, accepting the output of the standard Linux “objdump” program as its input. The program parses the objdump output and when it encounters a function, the algorithm checks that function signature against a blacklist. This blacklist contains regular expressions that match built in or global functions. If there is a match to the blacklist, the function is not considered in the totals, as these functions represent library elements which were not shuffled; C calls such as “printf” for example are exempt. If the function is not on the blacklist, then the statistics from the function are included in the data.

The benchmarks elected were compiled and tested on a HP ProLiant processor, 2.67GHz, with 160 GB installed, and a total of 24 cores available. Note however that while the multiple cores are used for faster compilation of the benchmarks the benchmarks themselves are not threaded and use only one CPU.

For technical reasons it is not possible to have a maximum LLVM block size setting that is less than three LLVM instructions. It thus seemed reasonable to test with the maximum block size ranging from three to ten, as via observation it did not appear that there were many LLVM blocks that were extremely large.

5.2 Performance Impact - Speed

To test the impact on execution speed, each of the benchmark programs was executed with a different maximum LLVM block size, ranging from three to ten. The execution runs were repeated 10 times and the number of seconds for each benchmark is the average of these times. (Note that benchmark 625 is three programs, and here the total execution time for all is shown. But 625 is broken out for the graphs concerning program size.) As expected, programs with more blocks execute more slowly, as the effects of the additional branch instructions and poorer cache performance due to a less optimum “principle of locality” have an impact. In figure 1 note the average speed (dashed line). The impact is not overly significant if the block size is limited to, say, six.

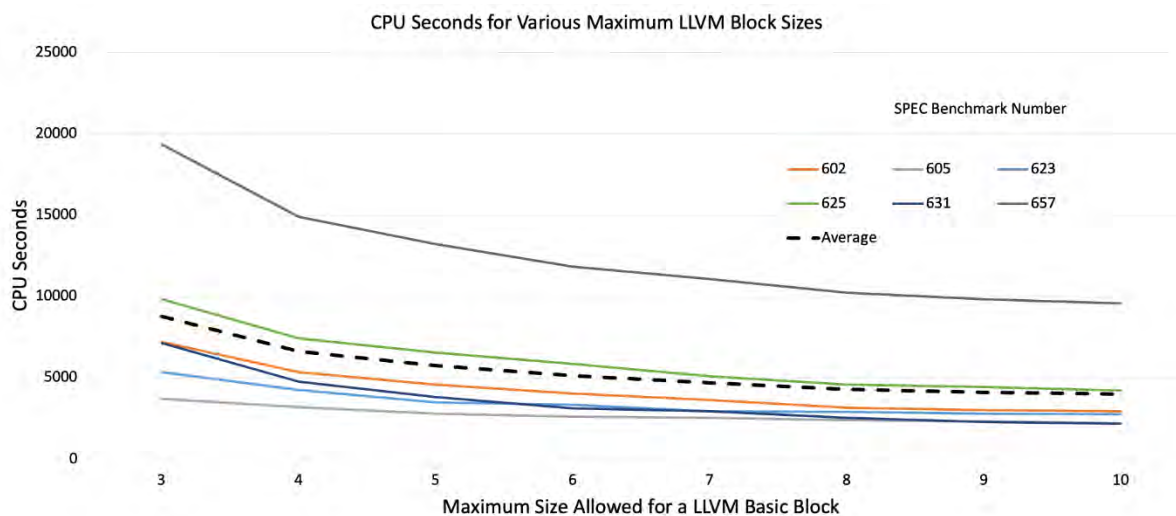


Figure 1: CPU Seconds Relative to Limited Maximum Block Size in LLVM Instructions

5.3 Performance Impact - Space

The authors next consider several effects in terms of program size. First, if one sets the maximum number of instructions in an LLVM block to N , the average number of instructions in a block will of course be less than N . Figure 2 below shows this impact. At lower values for N , three for example, the setting dominates because many or most blocks are larger than three instructions. But as the setting is increased, fewer and fewer blocks are as large as the setting, and the graph flattens. Eventually a larger and larger setting will have no impact once the setting is larger than any possible block in the program.

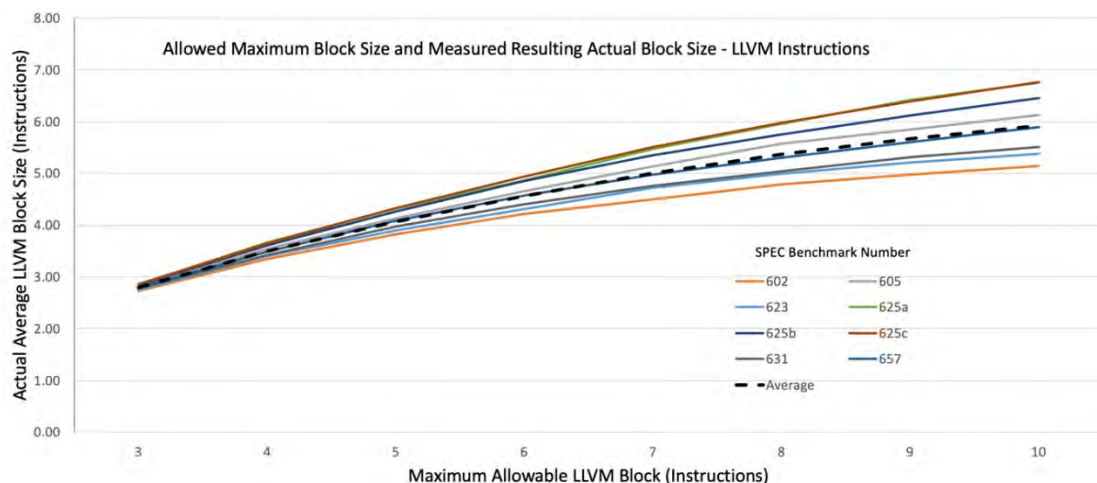


Figure 2: Average LLVM Instructions per Block as the Maximum Value is Increased

The next consideration is the relationship between LLVM block instruction count versus x86 block instruction counts. This is not a one-to-one relationship, because one LLVM instruction may produce more than one x86 instruction, or conversely two or more LLVM instructions may be converted into one complex x86 instruction. However, surprisingly the data indicate that after a point there is a close match between the average number of x86 instructions per block versus the original LLVM instructions per block. At a maximum block size of three there are 73% more x86 instructions than LLVM instructions – a high overhead. But at a maximum of 10 this overhead is down to just 5%.

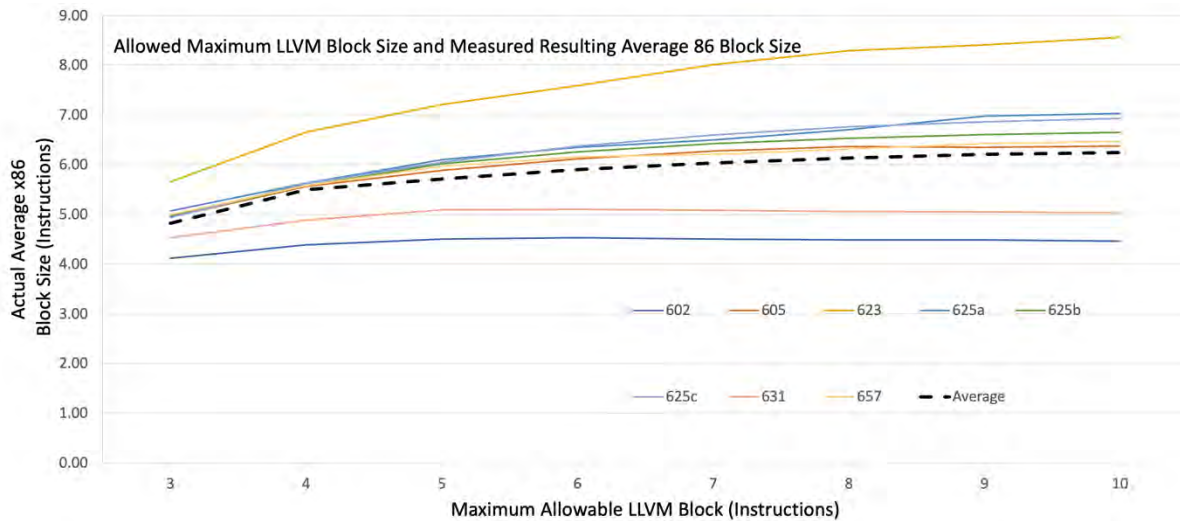


Figure 3: Allowed Maximum LLVM Block Size and Measured Resulting Average x86 Block Size

The final question is the number of actual machine code blocks that are generated based on different maximum LLVM settings. Again, this will not be a one-to-one mapping for several reasons. If after the blocks are shuffled, for example, two LLVM blocks end up in the proper sequence, the jump instruction from the first to the second block will be eliminated when the x86 code is generated. Recall that a motivation for smaller blocks is to increase the data carrying capacity of the fingerprint.

The number of x86 blocks generated from the compilation process, versus the maximum LLVM size, is depicted in the figure 4. The figure shows again that the maximum impact is with very small settings for the maximum number of instructions.

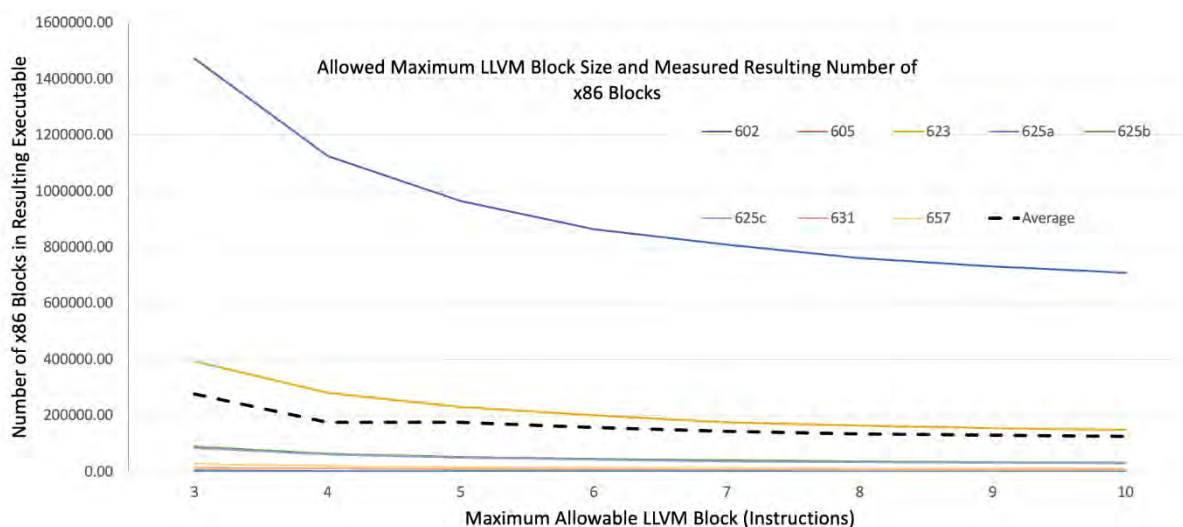


Figure 4: Allowed Maximum LLVM Block Size and Measured Resulting Number of x86 Blocks

As mentioned previously the data carrying capacity is dictated by the number of blocks. Recall that previously the example used was five blocks with a capacity of $5! = 120$, versus six blocks with a capacity of $6! = 720$ messages. The actual number of messages is less. Consider that in the process of making a permutation of blocks it is

possible that some of the blocks will be shuffled into their correct original ordering. When LLVM blocks are translated into x86 blocks, these adjacent blocks may be combined into one, removing the unnecessary jump instruction. One would expect the x86 block count to likely be smaller than the LLVM block count. Interestingly this is not the case, as shown in figure 5:

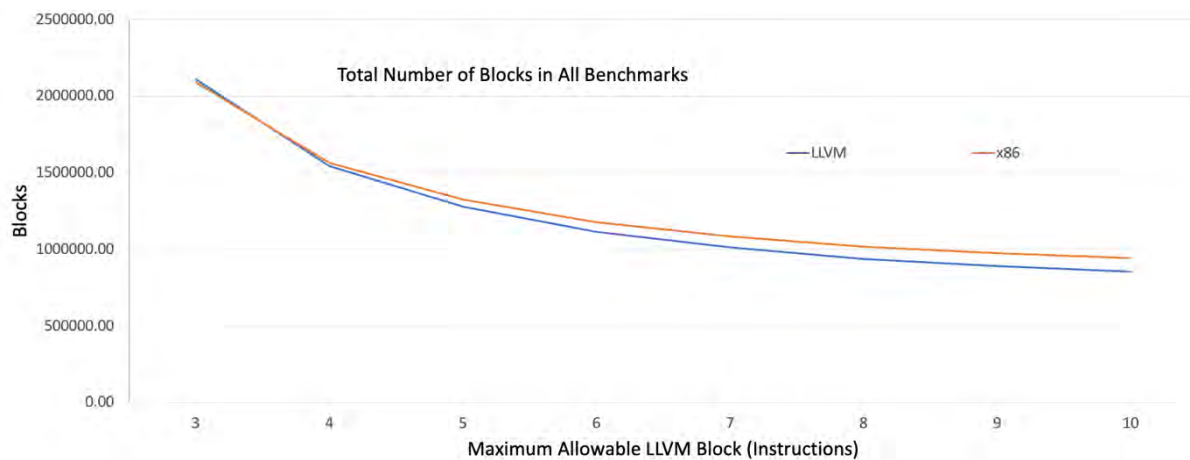


Figure 5: Total Number of Blocks in All Benchmarks – x86 and LLVM

We are investigating this discrepancy in more detail.

6. Conclusions

The use of block shuffling for software fingerprinting, using the LLVM infrastructure, has less of an impact in terms of size and speed as one might assume. However, there is a critical tradeoff that must be acknowledged. This tradeoff is related to the amount of information that can be “carried” (the number of permutations) versus the impact on the performance. This leads us to suggest that for fingerprinting purposes, one should select functions which are not frequently executed, but which have a higher degree of complexity. This greater complexity leads to a larger number of blocks, thus decreasing the need to split the blocks for fingerprinting purposes. Should the blocks need to be increased again, selecting an infrequently executed function will decrease any impact on the performance of the program.

In terms of execution speed, prior work (Mahoney 2021) indicated a speed overhead of approximately 1%. This was measured with no restriction on the block size. Our current work demonstrated a significant overhead when the maximum size of a block is severely restricted to a low number such as three or four instructions. In the case of limiting this block size to four, for example, the overhead is 66% (relative to a maximum of 10); but limiting the block size to six decreases this overhead to 29%. It thus would be possible to select a limiting factor based not only on the necessary data carrying capacity of a function (on the order of $N!$ for a function with N blocks) but also with an eye to limiting the performance overhead at the same time.

In terms of space overhead, our prior work reported results that indicated a space overhead of approximately 1%, but with no limitation on the block sizes as we investigate in this paper. Decreasing the allowable size of an LLVM block will increase the space overhead of the resulting program as there will be a larger number of jump (branch) instructions in the resulting executable. An examination of figures 1 and 2 indicates that as the instructions per block limit is increased, the average instructions per block matches closely; this is true for the resulting x86 block sizes as well. Figure 3 shows that there is typically a larger actual x86 instruction count than LLVM instruction count. For example, if an LLVM block is limited to three, the resulting average in terms of x86 is around five instructions. However, as figure 3 indicates, this is only extreme at the lower values for this maximum. Figure 4 shows that the number of blocks (related to the data capacity, recall) does not change significantly if the limitation on block size is over four. Finally, a comparison of the number of LLVM blocks versus the number of x86 blocks (figure 5) shows that the two track rather closely. This combination of facts tends to say that restricting the block size within the LLVM intermediate representation will of course lead to a larger number of blocks and thus a higher data rate, but that the restriction likely should not be less than five or six instructions per block.

Lastly, the best method for increasing the data rate without suffering a significant overhead penalty might be to fingerprint only certain functions which are rarely called or only occasionally executed – startup code for the program, or exception handlers, for example, where the data rate can be increased by lowering the block size but with little performance impact. Given that the number of fingerprint combinations is directly related to the number of blocks and that this grows as a factorial function, it may be that this approach is the best for software intellectual property protection via fingerprinting with block shuffling.

Acknowledgements

This work is based upon work funded by the National Science Foundation under the Secure and Trusted Computing (SaTC) grants CNS-1811560 and 1811578. The project is a collaborative effort between the University of Nebraska at Omaha (UNO) and the University of South Alabama (USA).

References

- Adve, V. and C. Lattner (2007) “2007 LLVM Developers’ Meeting (Video)”, [online], Available: <https://www.youtube.com/watch?v=FNtmemyeEHY&feature=youtu.be>.
- Alrehily, A., and Thayanathan, V. (2017) “Software Watermarking based on Return-Oriented Programming for Computer Security”, *International Journal of Computer Applications*, vol. 166, no. 8, pp. 21-28.
- Arboit, G. (2002) “A Method for Watermarking Java Programs via Opaque Predicates”, in *International Conference on Electronic Commerce Research (ICECR-5)* Montreal.
- Bento, L. M. S., Boccardo, D. R., Machado, R. C. S., d. Sá, V. G. P., and Szwarcfiter, J. L. (2019) “Full Characterization of a Class of Graphs Tailored for Software Watermarking”, *Algorithmica*, vol. 81, pp. 2899-2916.
- Colberg, C. and Thomborson, C. (1999) “Software watermarking: models and dynamic embeddings”, in *ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, San Antonio, Texas.
- Colberg, C. S. and Thomborson, C. (2002) “Watermarking, tamper-proofing, and obfuscation - tools for software protection”, *EEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 735-746.
- Colberg, C., Huntwork, A., Carter, E., and Townsend, G. M. (2004) “Graph Theoretic Software Watermarks: Implementation, Analysis, and Attacks”, *Information and Software Technology*, vol. 51, no. 1, pp. 192-207.
- Cousot, P., and Cousot, R. (2004) “An abstract interpretation-based framework for software watermarking”, *ACM SIGPLAN Notices*, pp. 173-185.
- Davidson, R. I., and Myhrvold, N. (1996) “Method and system for generating and auditing a signature for a computer program”. Patent 5,559,884.
- Fischer, R. A., and Yates, F. (1938) *Statistical tables for biological, agricultural and medical research* (3rd ed.) London: Oliver & Boyd.
- Hamilton, J., and Danicic, S. (2020) “A Survey Of Graph Based Software Watermarking”, [online], Available: <https://jameshamilton.eu/sites/default/files/GraphWatermarkingSurvey.pdf>
- Junod, P., Rinaldini, J., Wehrli, J., and Michielin, J. (2015) “Obfuscator-LLVM -- Software Protection for the Masses”, in *2015 IEEE/ACM 1st International Workshop on Software Protection*, Florence.
- Kamela, I., and Albluwib, Q. (2009) “A robust software watermarking for copyright protection”, *Computers and Security*, vol. 28, no. 6.
- Knuth, D. (1969) *Seminumerical algorithms. The Art of Computer Programming.*, Reading, MA: Addison-Wesley, pp. 139-140.
- LLVM (2021) “opt - LLVM optimizer”, [online], Available: <http://llvm.org/docs/CommandGuide/opt.html> and retrieved October 2021.
- Ma, H., Jia, C., Li, S., Zheng W., and Wu, D. (2019) “Xmark: Dynamic Software Watermarking Using Collatz Conjecture”, *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 11, pp. 2859-2874.
- Mahoney, W., Franco, J., Hoff, G., and McDonald, J. T. (2018) “Leave it to Weaver”, in *8th Software Security, Protection, and Reverse Engineering Workshop*, San Juan, Puerto Rico.
- Mahoney, W., Hoff, G., McDonald, J. T., Grispos G. (2021), “Software Fingerprinting in LLVM”, 16th International Conference on Cyber Warfare and Security.
- Mullins, J.A., McDonald, J. T., Mahoney, W., and Andel, T. (2020) “Evaluating Security of Executable Steganography for Digital Software Watermarking”, in *Cybersecurity Symposium*, Moscow, Idaho.
- Myles, G., and Colberg, C. S. (2006) “Software watermarking via opaque predicates: Implementation, analysis, and attacks”, *Electronic Commerce Research*, vol. 6, no. 2, pp. 155-171.
- Nagra, J., Thomborson C. D. and Colberg, C. S. (2002) “A Functional Taxonomy for Software Watermarking”, in *Twenty-Fifth Australasian Computer Science Conference (ACSC2002)* Melbourne, 2002.
- Palsberg, J., Krishnaswamy, S., Kwon, M., Ma, D., Shao Q., and Zhang, Y. (2000) “Experience with software watermarking”, in *Proceedings 16th Annual Computer Security Applications Conference (ACSAC'00)* New Orleans, LA.
- Preda, M. D., and Pasqua, M. (2017) “Exception Handling-Based Dynamic Software Watermarking”, *Electronic Notes in Theoretical Computer Science*, vol. 331, pp. 71-85.
- Qu G., and Potkonjak, M. (1998) “Analysis of watermarking techniques for graph coloring problem”, in *IEEE/ACM International Conference on Computer Aided Design*, San Joe.

- SPEC (Standard Performance Evaluation Corporation) (2017) [online], Available: <https://www.spec.org/cpu2017/>
- Stern, J. P., Hachez, G., Koeune, F., and Quisquater, J. J. (1999) "Robust Object Watermarking: Application to Code", in *Information Hiding; Lecture Notes in Computer Science volume 1768*, Berlin, Springer, pp. 368-378.
- Venkatesan, R., Vazirani, V., and Sinha, S. (2001) "A graph theoretic approach to software watermarking", in *4th International Information Hiding Workshop*, Pittsburgh.
- Wang, Y., Gong, D., Lu, B., Xiang, F., and Liu, F. (2018) "Exception Handling-Based Dynamic Software Watermarking", *IEEE Access*, vol. 6, pp. 8882-8889.

Taxonomy of Social Engineering Attacks: A Survey of Trends and Future Directions

Arianit Maraj¹ and William Butler²

¹Telecom of Kosovo, Technical Department, Republic of Kosovo

²Capitol Technology University, Laurel, Maryland, USA

arianit.maraj@kosovotelecom.com

whbutler@captechu.edu

Abstract: Hackers have many techniques available for breaching the security flaws of organizations. The human approach, called Social Engineering (SE), is probably the most difficult one to be dealt with. Social engineering is considered one of the most creative methods for gaining unauthorized access to information systems. This type of cyber threat does not require advanced technical knowledge because it relies mainly on human nature. Social engineers use different techniques, such as phishing, to manipulate people and cause significant damage to the organizations where they work. Therefore, organizations must raise the awareness of their users about social engineering attacks. Most organizations are putting all defense efforts into advanced technologies to prevent various threats. This is considered a wrong approach because employees of an organization use email, social networks, or other online sites as part of their work activities. Therefore, the prevention of attacks cannot be accomplished through advanced technologies alone, but the human aspect must also be studied. This paper comprehensively analyzes the existing literature in the taxonomy of social engineering attacks focusing on human aspects. It provides an overview of research opportunities that should be addressed and elaborated in future investigations.

Keywords: Social engineering, human factors, phishing, trust, perceptual, socio-emotional, education, awareness-raising, Social Engineering taxonomy

1. Introduction

With the development of new technologies, the number of devices connecting to the Internet is increasing exponentially. Today, there are approximately 46 billion Internet-connected devices (Juniper Research, 2016). With the enormous growth of devices connected to the Internet, the risk of misuse of sensitive data transmitted by these devices is increasing.

Various technologies have been developed to protect sensitive online data from cybercriminals. However, not all attacks can be detected through advanced technologies. Social engineering attacks require human interactivity, so protection against these attacks requires a different approach. Social Engineering (SE) attacks are becoming a major threat to organizations and individuals. Cybercriminals usually try to access sensitive data, such as passwords, bank account information, or access the organization's system to control the entire system (Airehrour, D. et al, 2018). These cybercriminals are using SE as a form of psychological manipulation of people into performing actions and disclosing confidential information. SE attacks are challenging to counter because they are mainly designed to manipulate human nature to disclose sensitive information online.

In general, security threats are divided into two types: technical and social engineering attacks. Technical threat means using advanced technologies to attack various individuals or organizations (Alharthi, D.N. et al., 2020, Applegate S. D., 2009). However, it has been proven that technical threats have not been effective in many cases so far. Therefore, hackers use SE techniques to bypass technical controls or advanced technologies implemented by psychologically manipulating users (Hadrnaghy C, 2010).

Social engineering attacks involve the psychological manipulation of users. Usually, they involve emails or other forms of communication, claiming in various false urgencies, fear, emotions, leading the victim to disclose sensitive data. Since SE involves a human element in each case, preventing these attacks is very complex, especially for organizations.

According to (Purplesec, 2021), 98% of all cyber-attacks rely on SE attacks. Therefore, it is imperative to take preventative measures to protect systems of particular importance and sensitive data. While cyber-attacks are a constant war, the only way to prevent them is to be aware of the dangers after every click on the Internet.

Through social engineering approaches, attackers can compromise many Internet-connected devices to create a Botnet (PDSC, 2020). As a result, criminals can carry out DoS (Denial of Service) (Maraj, A. et al., 2017), SQL

Injection (Maraj, A. et al., 2017), and attacks on various networks (Lepaja, S., et al, 2018) and technologies, such as IoT (Internet of Things), Web of Things (Baraković, S., et al., 2020, Maraj, A., et al., 2019), etc.

SE attacks are becoming more common against various organizations and becoming more sophisticated. Cybercriminals are creating intelligent methods to deceive the employees of an organization, so organizations must follow proactive approaches and models to defend themselves against these attacks. Since today, individual Internet users use the Internet to perform various services, such as online shopping or other services, attackers are focusing on attacks on individuals. This paper addresses and discusses the relationships and interplay of technology, human behavior, and procedural countermeasures to protect sensitive data.

The paper is organized as follows: after the introductory section, in Section 2, we provide a theoretical background regarding the taxonomy of social engineering attacks. Section 3 describes the research methodology. Section 4 gives an overview of the existing literature addressing the impact of human factors on SE attacks. The human factors in SE and taxonomy proposal are provided in Section 5, while Section 6 concludes the paper.

2. Theoretical background

2.1 Definition of social engineering

Social engineering is one of the most creative methods to gain unauthorized access and obtain sensitive information from individuals or organizations (Alsulami, M.H., et al, 2021). SE is a collection of different techniques used to manipulate people into disclosing confidential information. The term SE is usually applied to fraud involving information collection or access to a sensitive system. In most cases, the attacker never comes face to face with the victims, and the victims rarely realize they have been manipulated (Thapar, A., 2007). Of all the information security points, people have become the weakest point. Hackers take advantage of this fact by psychologically manipulating people to persuade them to disclose sensitive information or commit malicious acts. This is why social engineering security attacks are challenging to detect and prevent (Airehrour, D., et al., 2018). To prevent SE attacks, organizations and their employees and ordinary users should be aware of the defense mechanisms that can mitigate the risk of these attacks.

Organizations mainly implement advanced technologies to detect and prevent security attacks. However, security is considered strong as the system's weakest point. According to (Schneier B., 2015), people represent the weakest link in the security chain and are responsible for the failure of a specific system. Apart from the 'human factor,' the security chain comprises technical and physical links (Kassner, M., 2020).

2.2 Social engineering categories

Social engineering attacks are classified into two main categories; non-technical attacks (human-based) and technical SE attacks (Thapar, A., 2007).

2.3 Human-based technique for hacking

Social engineering uses human behavior to exploit systems and sensitive data. In human-based attacks, the attacker interacts directly with the human target to gather the desired information (Salahdine, F. et al., 2019). Although new technologies have been used to prevent security threats, human factors have been neglected. According to some recent research, it has been discovered that today there are terms that relate to SE and go beyond technology and are based on human error and social psychology (Peltier, T.R., 2006).

Human-based attacks involve the interaction of the attacker and the victim who possesses valuable information. An SE attacker initially collects sensitive information from the target, and then through the collected information, the attackers aim to build relations with the target to gain trust. The attacker tries to persuade the victim to perform the desired actions (e.g., disclosing confidential information). The attacker uses the information gathered to carry out attacks in the final stage. Social engineering attacks are undoubtedly one of the weakest links in cybersecurity since these attacks rely on human interactions, bypassing completely technical security mechanisms (Luo X., et al., 2011).

2.4 Technical-based attacks

Through technical-based attacks, cybercriminals try to access confidential information using computer software, email attachments, and websites (Luo X., et al., 2011). In technical-based hacking, cyber attackers use advanced

techniques to gain unauthorized access. However, it is difficult for hackers to successfully attack computer systems and networks using only technical means (Applegate S. D., 2009, Alharthi, D.N et al., 2020). About 50% to 75% of cyber security threats are due to misuse of information by humans, intentionally or unintentionally (Choi M., et al., 2013). Therefore, the main focus of this paper will be on human-based attacks.

3. Research Methodology

The main goal of this paper is to contribute to SE defense methods and countermeasures. To achieve that goal, we must first understand the taxonomy of SE attacks and the role of the human factor in SE attacks. We first collected the latest literature related to SE attacks and the impact of the human factors in preventing such attacks. Afterward, we survey and compare numerous studies contributing to this field. To simplify and summarize, the main objectives to achieve the primary goal of this paper are:

- To gain a deeper understanding of the SE attacks and factors for addressing such attacks.
- To organize and synthesize the existing research literature from the field;
- To identify gaps and issues needed to be addressed by the future research activities and suggest novel and open investigation directions in the subject field.

Our research methodology relies on guidelines given by authors in (Bandara, W., et al., 2013, Levy, Y et al., 2006), Figure 1. It consists of three phases. In this manuscript, we surveyed papers coming from both the technical sciences domain and the social sciences.

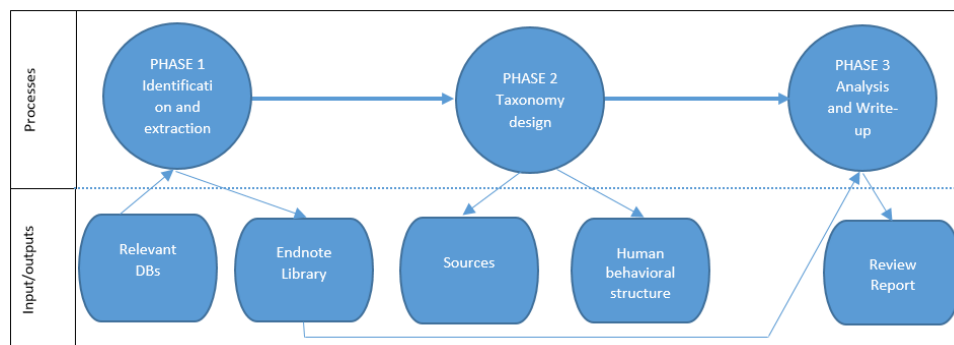


Figure 1: Research Methodology phases

The first phase includes identifying and reviewing the relevant literature regarding social engineering attacks and human factors influencing SE attacks. The second phase includes the taxonomy design. The third phase includes the analysis of the selected literature and the write-up process to identify the most common SE impacting human factors. The paper will summarize how human factors can play a crucial role in increasing overall security.

4. Literature review analysis

A wide range of research focuses on technical security attacks, while fewer researchers have focused on social engineering attacks and human behavior. This section discusses research efforts related to technical and non-technical social engineering attacks.

(Ivaturi, K. et al., 2011) discussed the different types of social engineering attacks. They used the latest literature to compare and describe Social engineering methods and the factors. This manuscript presented a taxonomy approach considering human factors in social engineering attacks. The authors analyzed different types of attacks that might occur in social networking platforms, such as; person-person, person-person via media, person-person using voice, and person-person via video.

The human factors that affect social engineering attacks are considered in (Saridakis, G., et al., 2016). The authors in this paper explore how behavioral patterns impact social networks, personal characteristics, and technical efficacy of users, impacting the risk of online victimization. As human factors, they have analyzed the role of education and self-confidence in preventing SE attacks in social networking platforms, particularly in Facebook, LinkedIn, google+, and blogger. They found out that the overall intensity of social media usage alone may not increase the risk of becoming a victim of cybercrime.

Human factors impacting susceptibility to phishing attacks were studied in (Iuga, C., et al., 2016). The authors considered the relationship between demographic characteristics of individuals and their ability to detect phishing attacks and time-related factors. They gathered cursor movement data to provide additional insight. In the end, they found that gender and the years of PC usage impact the detection of phishing attacks. According to this study, many people are still at high risk of falling victim to phishing attacks. They suggest that we have to consider using a different combination of automated tools, training, and awareness-raising campaigns to prevent such attacks. The influence of demographic factors on SE attacks are also addressed in (Tynes, B.M. et al., 2014, Duggan, M., et al., 2012, and (Peltier, T. 2006).

The trust factors of social engineering attacks are discussed in (Kano, Y. et al., 2021). The authors experimented with 35 participants using pseudo-social networking service applications facing SE attacks. Trust factors are considered in (Alseadoon IMA, 2014) and (Workman M, 2007).

The authors in (Orgill, G.L., 2004) present the results of the SE audit made without any notice within one organization where data security and privacy are the primary concerns. The privacy awareness impact on SE is discussed in (Krombholz, K., 2015) and (Orgill, G.L., 2004).

Authors in (Aldawood, H. et al., 2018) have demonstrated how education programs can increase awareness and reduce cyber security incidents. Literature such as (Chen Y. et al., 2015, Parsons K., et al., 2014) analyzed factors related to education and awareness-raising. To prevent social engineering attacks, many scholars suggest that there is a need to increase the level of education and awareness and implement training programs for users and employees (Parsons K. et al., 2014, Al-Daeef M. M., et al., 2017, Amankwa E., et al., 2014, Arachchilage N. A. et al., 2014, Ashford W., 2016, Eyong B. K., 2014, Lebek B., 2013, Maraj, A., et al., 2021, Maraj, A., et al., 2020, Maraj, A., et al., 2021)

The impact of human perceptual factors in social engineering attacks is considered in (Kearney, W.D. et al., 2016). According to their analysis, organizations could influence information security regarding the perception of a specific threat. User perceptions related to potential threats are also considered in (Staggs, J. et al., 2014, Vance, A., et al., 2021). Through education, training, and awareness-raising programs, user awareness can be increased (D'Arcy, J. et al., 2009).

Table 1: Literature review analysis

Ref.	Clear aim	Research methodology	Human factors	Technical factors	Types of attacks considered	Platform	Findings	Impact
Ivaturi, K. et al., 2011	The papers aim is to classify social engineering attacks through taxonomy	The authors of this manuscript used the latest literature to compare and describe Social engineering methods and the factors	The authors considered human factors	Technical factors not included	Person-person, person-person via media, person-person using voice, person-person via video	Social networking	The authors brought some clarity to the different types of social engineering attacks through their proposed taxonomy approach	This paper has a significant impact on analyzing some of the most popular SE human-based attacks.
Saridakis, G., et al., 2016	The authors tend to explore how behavioral patterns on social networks, personal characteristics, and the technical efficacy of users affect the risk of facing	Survey	They considered the education factor	Technical factors not considered	In general, all SE-related attacks	Social networking platform; Facebook, LinkedIn, google+, blogger	The authors of this study found that the overall intensity of social media usage may not increase the risk of becoming a victim of cybercrime.	This paper has some limitations, but still, it has a significant impact since it opens some future empirical work in the field.

Ref.	Clear aim	Research methodology	Human factors	Technical factors	Types of attacks considered	Platform	Findings	Impact
	online victimization.							
Peltier, T., 2006	The authors examine how people, government agencies, military organizations, and companies have been duped into giving sensitive information.	The authors used the latest literature to compare and describe Social engineering attacks; technology and non-technology-based attacks.	The authors considered gender factors and education	The authors considered the technology-based attacks; windows pop-ups, mail attachments, websites	Impersonation, in-person	Internet	Employee education is the key	This paper has a significant impact on social engineering attacks
Kano, Y. et al., 2021	This paper explores the trust factors of social engineering attacks on SNSs.	The authors experimented with 35 participants using a pseudo-social-networking-service application facing a social engineering attack.	The authors considered trust as a human factor	Not considered	All SE-related attacks in Social networking services	Social networking services	They found that the profile details of the attacker and the content of the post do not affect trust.	This paper has a significant impact since it addresses the risks in social networks.
Aldawood, H. et al., 2018	The paper details how security education programs can increase user/employee awareness and reduce cyber security incidents.	Literature review	The authors considered the education and awareness	Not considered	Phishing	Social media, jobs portals, web page	The authors found that implementing information security education and awareness programs can be effective to increase user awareness for preventing cyber-crimes	This paper is important because the authors have analyzed many recent studies in this field.
Kearney, W.D., et al., 2014	The authors try to explore how fear and self-confidence influence both intention to respond and actual response to a Phishing attack.	This study collected data from multiple sources.	Self-confidence and fear	Not considered	Phishing	Social media, Internet, Email	They found that as more knowledge is gained regarding how fear of providing login credentials can affect responses to phishing attacks and how self-confidence can affect the relationship between fear and intention to respond.	This research provides valuable information to assist in preventing the use of social engineering to obtain sensitive information.

The literature review results in this domain are depicted in Table 1. We found some important papers that address the impact of human factors in Social Engineering attacks. We selected six papers addressing human factors in different platforms, mainly on social networking platforms.

5. Impact of human factors in Social Engineering attacks and taxonomy proposal

To summarize the literature review analysis on human factors' impact on SE, several findings are depicted in Table 2. In Table 2, we selected eighteen papers that address the main human factors influencing the SE. We have noticed from this analysis that none of these papers include most of the human factors that affect SE. Therefore, this concludes that there is a need for studies that address the full range of impacting human factors in SE. Table 2 indicates that several papers address only demographic and education factors (Iuga, C. et al., 2016). Other papers address only demographic factors (Tynes, B.M. et al., 2014, Duggan, M., et al., 2012). Privacy awareness is considered by (Duggan, M. et al., 2004) and (Krombholz, K. et al., 2015). Education, Awareness raising, and Training programs are beneficial to reduce the SE risk in organizations. These factors are considered in (Iuga, C. et al., 2016, Saridakis, G., et al., 2016, Peltier, T. 2006, Luo X et al., 2011, Orgill, G.L., et al., 2004, Aldawood, H. et al., 2018). However, none of these papers (Table 2) addressed the full range of these human behavioral factors.

Table 2: Review of Human factors in Social Engineering attacks

Ref/Human factors	Users demographics; Age, gender	Socio-emotional perspective	Self-confidence	Trust	Perceptual	Privacy awareness	Education/Awareness/Computer skills
Iuga, C., et al., 2016	✓						✓
Saridakis, G., et al., 2016			✓				✓
Tynes, B.M. et al., 2014	✓						
Duggan, M., et al., 2012	✓						
Peltier, T., 2006	✓						✓
Luo X, et al., 2011							✓
Kearney, W.D., et al., 2016					✓		
Alseadoon IMA, 2014		✓		✓			
Workman M, 2007				✓			
Staggs, J., et al., 2014					✓		
Orgill, G.L., et al., 2004						✓	
Krombholz, K., et al., 2015						✓	
Kano, Y. et al., 2021				✓			
Orgill, G.L., et al., 2004						✓	✓
Aldawood, H. et al., 2018							✓
Pimentel, A. et al., 2021		✓					
Kearney, W.D., et al., 2014				✓			
House, D. et al., 2020			✓				

To reduce users' vulnerability to social engineering attacks, we first need to understand users' behavior and consider the impacting human factors.

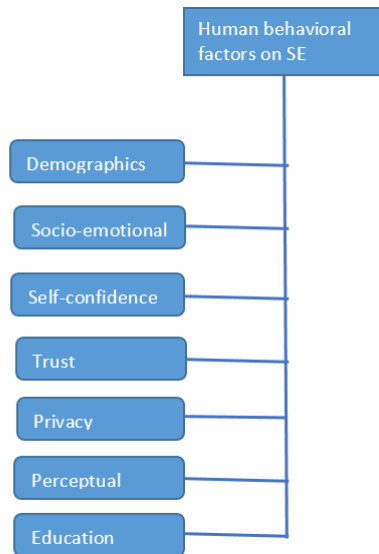


Figure 2: Proposed taxonomy

The literature review revealed no previous studies investigating the full range of human impacting factors involved in human behavior. Most of the research focused on only one type of behavior. This study proposes a taxonomy, which will involve most of the human factors influencing SE attacks; users demographics, socio-emotional perspective, confidence, trust, perceptual, privacy, and education (Figure 2). We consider deficient any other analysis that does not address the full range of factors. Therefore, we think that there should be a full range of parameters.

6. Conclusions

This paper addresses and discusses the human factors which impact SE attacks. The ultimate goal is to contribute to security improvement by considering the full range of human behavioral factors on SE. The contribution of this review paper is three-fold. Firstly, we have proposed the SE taxonomy, which will address the full range of human factors impacting SE attacks. Secondly, the literature analysis and comparison have been conducted based on that taxonomy. The findings indicate that the full range of human factors have yet to be fully addressed so far. Thirdly, we have identified the gaps in human factors impacting the SE attacks. It is essential to discover how considering a full range of human factors will improve overall security, thus minimizing SE attacks and their impacts. We are convinced that the likelihood of SE attacks can be mitigated if the employees of the organizations understand their essential role in the overall strategy for SE defense. The dilemma is that organizations and their employees are at a disadvantage because they rarely face the inevitable exposure to SE attacks. Future research should focus on the least studied human factors such as Socio-emotional perspective and ones Self-confidence for example. In addition future studies should also focus on how the application of Artificial Intelligence and Machine Learning could disrupt or minimize the effects of SE attacks.

References

- Airehrour, D., Nair, N.V. and Madanian, S. (2018) 'Social engineering attacks and countermeasures in the New Zealand banking system: advancing a user-reflective mitigation model', *Information*, Vol. 9, No. 5, p.110. doi:10.3390/info9050110
- Al-Daeef M. M., N. Basir, and M. M. Saudi, "Security awareness training: A review," in *Lecture Notes in Engineering and Computer Science*, 2017, vol. 2229, pp. 446–451.
- Aldawood, H. and Skinner, G., 2018, December. Educating and raising awareness on cyber security social engineering: A literature review. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 62-68). IEEE.
- Alharthi, D.N. and Regan, A.C., 2020, July. Social engineering defense mechanisms: A taxonomy and a survey of employees' awareness level. In *Science and Information Conference* (pp. 521-541). Springer, Cham.
- Alharthi, D.N., and Regan, A.C., 2020, July. Social engineering defense mechanisms: A taxonomy and a survey of employees' awareness level. In *Science and Information Conference* (pp. 521-541). Springer, Cham.
- Alseadoon IMA (2014) The impact of users' characteristics on their ability to detect phishing emails. Doctoral Thesis. Queensland University of Technology. <https://eprints.qut.edu.au/72873/>

- Alsulami, M.H., Alharbi, F.D., Almutairi, H.M., Almutairi, B.S., Alotaibi, M.M., Alanzi, M.E., Alotaibi, K.G. and Alharthi, S.S., 2021. Measuring Awareness of Social Engineering in the Educational Sector in the Kingdom of Saudi Arabia. *Information*, 12(5), p.208.
- Amankwa E., M. Looock, and E. Kritzing, "A conceptual analysis of information security education, information security training, and information security awareness definitions," in *Proc. 9th Int. Conf. Internet Technology and Secured Transactions (ICITST' 14)*, London, UK, 2014, pp. 248–252.
- Applegate S. D., Social engineering: hacking the wetware!, *Information Security Journal: A Global Perspective* 18 (1) (2009) 40{46.
- Arachchilage N. A. G. and S. Love, "Security awareness of computer users: a phishing threat avoidance perspective," *Comput. Human Behavior*, vol. 38, pp. 304–312, Sep. 2014.
- Ashford W., "Lack of cyber security awareness putting UK organizations at risk," *ComputerWeekly.com*, Mar. 2016.
- Bandara, W., Mikson, S.; Fieli, E. A Systematic, Tool-supported Method for Conducting Literature Reviews in Information Systems. In *Proceedings of the 19th European Conference on Information Systems (ECIS)*, Helsinki, Finland, 9–11 June 2013. 26.
- Baraković, S., Baraković Husić, J., Maraj, D., Maraj, A., Krejcar, O., Maresova, P. and Melero, F.J., 2020. Quality of life, quality of experience, and security perception in web of things: An overview of research opportunities. *Electronics*, 9(4), p.700.
- C. Hadnagy, *Social engineering: The art of human hacking*, John Wiley & Sons, 2010. 3. A. Berg,
- Chen Y., K. Ramamurthy, and K.–W. Wen, "Impacts of comprehensive information security programs on information security culture," *J. Comput. Inform. Syst.*, vol. 55, no. 3, pp. 11–19, Dec. 2015.
- Choi M., Y. Levy, A. Hovav, The role of user computer self-efficacy, cybersecurity countermeasures awareness, and cybersecurity skills influence on computer misuse, in: *Proceedings of the Pre-International Conference of Information Systems (ICIS) SIGSEC(Workshop on Information Security and Privacy (WISP)*, 2013.
- D'Arcy, J., Hovav, A., and Galletta, D. (2009). User Awareness of Security Countermeasures and its Impact on Information Systems Misuse: A Deterrence Approach, *Information Systems Research*, 20(1), 79-98.
- Duggan, M., Brenner, J., 2012. The Demographics of Social Media Users — 2012. *Pew Research Center Internet & American Life Project* (Available at: <http://www.lateledipenelope.it/public/513cbff2daf54.pdf> [Accessed on October 2021]).
- Eyong B. K., "Recommendations for information security awareness training for college students," *Inform. Manage. & Comput. Security*, vol. 22, no. 1, pp. 115–126, 2014.
- House, D. and Raja, M.K., 2020. Phishing: message appraisal and the exploration of fear and self-confidence. *Behavior & Information Technology*, 39(11), pp.1204-1224.
- Iuga, C., Nurse, J.R. and Erola, A., 2016. Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-centric Computing and Information Sciences*, 6(1), pp.1-20.
- Ivaturi, K. and Janczewski, L., 2011, June. A taxonomy for social engineering attacks. In *International Conference on Information Resources Management* (pp. 1-12). Centre for Information Technology, Organizations, and People.
- Juniper Research, Internet of Things Connected Devices to triple by 2021, available online at: <https://www.comparethecloud.net/news/internet-of-things-connected-devices-to-triple-by-2021-reaching-over-46-billion-units/>, accessed on December 2021
- Kano, Y. and Nakajima, T., 2021, March. Trust Factors of Social Engineering Attacks on Social Networking Services. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* (pp. 25-28). IEEE.
- Kassner, M., 2020, Cybersecurity pros: Are humans really the weakest link? Available at: <https://www.techrepublic.com/article/cybersecurity-pros-are-humans-really-the-weakest-link/> , accessed on November 2021
- Kearney, W.D., and Kruger, H.A., 2014, August. Considering the influence of human trust in practical social engineering exercises. In *2014 Information Security for South Africa* (pp. 1-6). IEEE.
- Kearney, W.D., and Kruger, H.A., 2016. Can perceptual differences account for enigmatic information security behavior in an organization?. *computers & security*, 61, pp.46-58.
- Krombholz, K., Hobel, H., Huber, M. and Weippl, E., 2015. Advanced social engineering attacks. *Journal of Information Security and applications*, 22, pp.113-122.
- Lebek B., J. Uffen, M. H. Breitner, M. Neumann, and B. Hohler, "Employees' information security awareness and behavior: A literature review," in *Proc. 46th Hawaii Int. Conf. System Sciences (HICSS' 13)*, Wailea, HI, 2013, pp. 2978–2987.
- Lepaja, S., Maraj, A., Efendi, I. and Berzati, S., 2018, June. The impact of the security mechanisms in the throughput of the WLAN networks. In *2018 7th Mediterranean Conference on Embedded Computing (MECO)* (pp. 1-5). IEEE.
- Levy, Y.; Ellis, T.J. A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Sci. Int. J. Emerg. Transdiscipl.* 2006, 9, 181–212.
- Luo X, Brody R, Seazzu A, Burd S. Social engineering: The neglected human factor for information security management. *Information Resources Management Journal (IRMJ)*. 2011 Jul 1; 24(3):1-8.
- Maraj, A., Atkinson, T. and Silaghi, M.C., 2019, May. Learning Strategies for Resisting Power Attacks on Wi-Fi Direct Group Formation. In *The Thirty-Second International Flairs Conference*.
- Maraj, A., Jakupi, G., Rogova, E. and Grajčevci, X., 2017, June. Testing of network security systems through DoS attacks. In *2017 6th Mediterranean Conference on Embedded Computing (MECO)* (pp. 1-6). IEEE.
- Maraj, A., Rogova, E. and Jakupi, G., 2020. Testing of network security systems through DoS, SQL injection, reverse TCP, and social engineering attacks. *International Journal of Grid and Utility Computing*, 11(1), pp.115-133.

- Maraj, A., Rogova, E., Jakupi, G. and Grajqevci, X., 2017, October. Testing techniques and analysis of SQL injection attacks. In 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA) (pp. 55-59). IEEE.
- Maraj, A., Sutherland, C. and Butler, W., 2021, June. Studying the Challenges and Factors Encouraging Girls in Cybersecurity: A Case Study. In ECCWS 2021 20th European Conference on Cyber Warfare and Security (p. 269). Academic Conferences Inter Ltd.
- Maraj, A., Sutherland, C. and Butler, W., 2021, June. The Challenges to Cybersecurity Education in Developing Countries: A Case Study of Kosovo. In ECCWS 2021 20th European Conference on Cyber Warfare and Security (p. 260). Academic Conferences Inter Ltd.
- Orgill, G.L., Romney, G.W., Bailey, M.G. and Orgill, P.M., 2004, October. The urgency for effective user privacy-education to counter social engineering attacks on secure computer systems. In Proceedings of the 5th conference on Information technology education (pp. 177-181).
- Parsons K., A. McCormac, M. Butavicius, M. Pattinson, and C. Jerram, "Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q)," *Comput. & security*, vol. 42, pp. 165–176, May 2014.
- PDSC, 2020, Denial Of Service (DOS) Attack/PDSC Advice Guide, available online at: <https://www.policedsc.com/security-advice/internet-of-things/denial-of-service-attack>, accessed on December 2021
- Peltier, T. (2006). Social Engineering: Concepts and Solutions. *Information System Security*, 15(5), 13–21. doi:10.1201/1086.106589 8X/46353.15.4.20060901/95427.3
- Pimentel, A. and Steinmetz, K.F., 2021. Enacting social engineering: the emotional experience of information security deception. *Crime, Law and Social Change*, pp.1-21.
- Purplesec, 2021 Cyber Security Statistics, available online at: <https://purplesec.us/resources/cyber-security-statistics/>, accessed on November 2021
- S. D. Applegate, Social engineering: hacking the wetware!, *Information Security Journal: A Global Perspective* 18 (1) (2009) 40–46.
- Salahdine, F. and Kaabouch, N., 2019. Social engineering attacks: A survey. *Future Internet*, 11(4), p.89.
- Saridakis, G., Benson, V., Ezingard, J.N. and Tennakoon, H., 2016. Individual information security, user behavior and cyber victimization: An empirical study of social networking users. *Technological Forecasting and Social Change*, 102, pp.320-330.
- Schneier, B., 2015. *Secrets and lies: digital security in a networked world*. John Wiley & Sons.
- Staggs, J., Beyer, R., Mol, M., Fisher, M., Brummel, B. and Hale, J., 2014, June. A perceptual taxonomy of contextual cues for cyber trust. In *Journal of The Colloquium for Information Systems Security Education* (Vol. 2, No. 1, pp. 10-10).
- Thapar, A., 2007. Social engineering: An attack vector most intricate to tackle. *CISSP: Infosec Writers*.
- Tynes, B.M. and Mitchell, K.J., 2014. Black youth beyond the digital divide: Age and gender differences in Internet use, communication patterns, and victimization experiences. *Journal of Black Psychology*, 40(3), pp.291-307.
- Vance, A., Siponen, M., & Pahlila, S. (2012). Motivating IS security compliance: Insights from habit and protection motivation theory. *Information & Management*, 49(3), 190-198.
- Workman M (2007) Gaining access with social engineering: an empirical study of the threat. *Inf Syst Secur* 16(6):315–331. <https://doi.org/10.1080/10658980701788165>

Blockchain Technology for Addressing Privacy and Security Issues in Cloud Computing

Pardis Moslemzadeh Tehrani¹, Gabriele Kotsis² and Andasmara Rizky Pranata¹

¹Faculty of Law, University of Malaya, Kuala Lumpur, Malaysia

²Head of the Telecooperation Department, Johannes Kepler Universität, Linz, Austria

pardismoslemzadeh@um.edu.my

Gabriele.kotsis@jku.at

andasmaraizky@gmail.com

Abstract: Blockchain technology is a recent and important financial technology that has completely transformed business transactions. The adoption of cloud computing in all IT environments due to its efficiency and availability has increased dramatically. Despite attempts to address privacy and security issues, confusion remains in the cloud environment. It is said that blockchains are a promising solution for many distributed applications and have the potential to overcome issues pertaining to the centralized system. The decentralized nature of blockchains provides a new forms of distributed software architectures where users can reach agreement consensually on the shared system without relying on a central integration point. Nonetheless, there are some challenges which need to be evaluated. Accordingly, this study explores how blockchains can curb cloud computing issues from both legal and technical aspects. This article presents the use of blockchain in current cloud storage applications and discusses how blockchain technology can be used to resolve security and privacy challenges. The use of blockchains as a solution for each issue will be proposed by explaining how they can overcome the shortcomings of cloud computing.

Keywords: Blockchain, Cloud Computing, Privacy, Security, Contractual Terms

1. Introduction

Cloud computing has been entrenched in our society due to its various advantages such as the storing of large amounts of data, accessibility, performance of intensive computational operations, etc. It represents an everything-as-a-service business model offering outsourcing services in an economic, scalable and flexible manner that is affordable and attractive to small businesses and individual consumers. Two important types of outsourcing services in cloud namely, storage and computation are widely available for users; however, data security and privacy issues remain of primary concern in the use of outsourcing services. (Zhang & Deng 2018)

The rapid development of cloud computing has raised many concerns over its usage due to the amount of risk or uncertainty inherently embedded in the layers of the protection in cloud storage. As the cloud deals with information that may involve all aspects of personal life tight security needs to be in place to protect individual privacy. Specific requirements on personal protection created by new regulations - particularly General Data Protection Regulation (GDPR) - challenge the effectiveness of the centralized computing model available on public cloud infrastructures. (Stanciu 2017)

The information being collected on an unprecedented scale and big data analysis in this big data era, led to a burgeoning demand for processing big data in a secure way with low cost and high efficiency. Cloud computing services have thus emerged and became popular for big data processing. Various and flexible approaches are offered by cloud service providers to data owners for managing their data sets that are stored remotely in the cloud. However, by nature, cloud storage is not secure. (Liu 2017) They are highly centralized architectures which suffer various technical limitations such as from cyber-attacks and single point of failure. In cloud computing, the data is stored in a centralized AI application which may lead to the possibility of data tampering and breaches as data can be subject to hacking and manipulation due to it being managed and stored in a centralized manner. (Salah et al. 2019)

According to Mungwe (2017), "The two main issues that exist with security and privacy aspects of cloud computing includes loss of control over data and total dependence on the cloud computing provider. (Mungwe 2017) The Council of European Professional Informatics Societies (CEPIS) explains these two issues more clearly: (Höbl 2017) "First, the data could be compromised by the cloud service provider (CSP) as there are limited ways to confirm the data transparency inside of the CSP. After all, some clouds have functions that enable CSP to take a sneak peek at the data stored in their server. Second, the connection between the server and the user may not be protected, hence no guarantee that all attempt at cybercrime will be repelled. Finally, it will be hard to

delete data permanently. It means that there is no guarantee that the master data that is deleted at the server is totally deleted, or even the master data anymore". CEPIS then notes that "the age of a 'Guaranteed complete deletion of data', if it ever existed, has passed." Typically, cloud users have no control over the cloud storage servers being used, thus giving rise to risks such as data confidentiality, integrity, availability, privacy, cybersecurity, energy consumption, and scalability. (Maroufi et al. 2019)

On the other hand, blockchain is a type of distributed electronic database, a ledger, which can hold any information (e.g., user data records, critical events information, banking transactions or device service history) and sets rules on how this information can be updated. Blockchain technology is used to organize a market network where users can monetize their computing power, applications, and datasets. (Bonhomme et al. 2020). It provides on-demand access to cloud computing resources to support compute-intensive applications in fields such as AI, big data, healthcare, rendering, or FinTech. Transactions in blockchains via a distributed ledger are verified by the consensus of a majority spread throughout the blockchain network. It allows non-trusting members to interact with each other without an intermediary as well as enable all participants to have their own copy of the ledger. Any minimal changes to the ledger will be reflected in all copies.

Despite attempts to address privacy and security issues, confusion remains in the cloud environment. (Tehrani et al 2017) It is said that blockchains are a promising solution for many distributed applications and have the potential to overcome issues pertaining to the centralized system. Nonetheless, there are some challenges which need to be evaluated. Accordingly, this study explores how blockchains can curb cloud computing issues from both legal and technical aspects. This article presents the use of blockchain in current cloud storage applications and discusses how blockchain technology can be used to resolve security and privacy challenges. The first part discusses the main privacy concerns of cloud computing in general and how blockchains provide a solution to those concerns. It then identifies other issues that derive from privacy concerns. The use of blockchains as a solution for each issue will be proposed by explaining how they can overcome the shortcomings of cloud computing.

2. Legal Issues Relating to Privacy and Security in Cloud Computing

Data privacy is a major issue in this era of networking system where many users share the same physical storage facilities. Cloud computing is normally used by many networking applications such as IoT, healthcare and smart grids to process and store massively large amounts of data. In fact, a critical requirement for most of these applications that are involved with personal information is privacy. However, the use of multiple clouds and internetworking among them has intensified the issue of privacy of the data. Under a privacy service the user has the right to control and set rules for the data and resources accessed by the network. The data and resource owner can control disclosure of their information, and this is generally done by enabling users to define their access control list (ACL). The following section discusses the challenges currently faced in providing privacy services in cloud storage.

Privacy and security are major issues in cloud applications that do not as yet have industry-wide solutions. The absence of cloud computing standards points to the need for common cloud security, data privacy and ownership policies due to the different approaches taken by service providers. As mentioned previously, cloud providers aim to ensure compliance with the laws and regulations governing the use of the internet by including policies protecting end users. The GDPR, for instance, accords the utmost importance to privacy and end users' data especially when it comes to sensitive information that may affect personal reputations. (Akbar 2018)

A major issue in this respect is the relationship between service providers and end users. Contrary to the common belief that data stored in the cloud belongs to the service provider, it should be remembered that it is still owned by the end user although the provider may own the software. Thus, it is understood that service providers should relinquish power to end users under the terms of the contract since the latter have no other option except to use the cloud due to its convenience and ubiquitous nature. One of the main grievances of end-users is the disowning of liability by cloud providers through the use of a clickwrap agreement. Such a contract in cloud computing upsets the balance between the parties to the agreement. The clickwrap agreement includes terms and conditions which are often unread or overlooked by end users when agreeing to the contract as it is provided separately from the acceptance page.

Another related issue is the presence of internal and external attackers in cloud storage infrastructures. (Privacy in Cloud Computing 2012) Internal attackers are those employed by the cloud service provider, customer, or other third-party provider organization supporting the operations of a cloud service. They have the opportunity to gain further access or support any third party in executing attacks against the confidentiality, integrity and availability of information within the cloud service. (TagElsir et al. 2015) They are deemed attackers as they can access the end user's personal data for their own benefit or profit. Most service providers scan the personal data of their customers in order to place targeted ads based on the information in the data. Moreover, more complex, data and statistics are recorded, bundled and analysed (data mining) for user profile marketing as a means to anticipate future purchases by users. The government is also considered an attacker of personal data. Users need to provide personal information which then enables the government to have massive amounts of personal data that should not be made public. Here, individuals have little or no control over storage and access to their information and this place the privacy of their data at risk. (Salman et al. 2019)

Various articles have discussed how technology size or privacy laws can help check the excessive influence of governments. (Adrian 2013) In *Microsoft Corp v United States*, (Microsoft Corp. v. United States 2016) the court held that the government cannot compel internet service providers to turn over stored data without a warrant. This is to prevent any unwarranted interception that deviates from grounds of national security. In fact, the court did not address the issue of whether the data resides in a particular country, and it was limited to extra-territorial requirements. Furthermore, in the interest of national security such as in combatting terrorism, governments have extensive legal avenues to access private information stored in the cloud. Essentially, cloud providers might be located in a different country from users which could lead to privacy issues if the cloud computing contract did not provide additional terms stating the level of protection and method of handling the data. (OECD 2014).

Another possible cause leading to a breach in privacy is the leakage in the cloud computing process in regard to the agreement between the processor and sub processor. Where the processor's job is delegated to a sub processor there is major concern over the status of the cloud provider as to whether it rests with the data controller or the data processor. Many providers have chosen to disavow being bound by liabilities that they should carry according to the role they play. Thus, users are the ones who normally bear the consequences if the processor and sub processor did not have a proper contract between themselves. It is important to determine whether a party is controller or processor when processing personal data, as under GDPR both have distinct obligations. (Salmon et al. 2019)

One who uploads data to a cloud computing environment is a controller under a cloud computing system while the operator is the processor. This is a key area in which blockchain and cloud systems differ. As many blockchain systems are operated by all users in a peer-to-peer network environment, it is difficult to define whether users are controllers or processors. It is necessary to consider the extent to which the different participants in the blockchain network are controllers based on their respective activities. Normally, participants who submit personal data to a blockchain are more likely count as controller based according to the GDPR since they determine the details of processing. The nodes, on the other hand, are more likely to be processors as they simply facilitate the blockchain network's operations by merely processing personal data. However, as not all blockchain systems operate in the same way, and there can be different types of participants carrying out various activities, this determination is not definitive and precise. As a representative technology on the anonymity aspect, blockchains can be upgraded to a convenient service that provides stronger security if combined with cloud storage.

2.1 Blockchain as a solution for privacy services

Blockchain technology can resolve some of the issues discussed in the previous section by providing decentralized end-to-end data privacy guarantees. In doing so, it can provide data ownership solutions and dynamically change access rights when required. An ideal blockchain-based data privacy system allows the owner to define the desired ACL through a smart contract that will publish the ACL and the data as blockchain transactions. Thus, many platforms such as Facebook or Google will not own the data as in the past. Instead, they will be a part of the blockchain network, processing data only when the ACL allows them via a permissionless blockchain where policies defining the data access are either based on smart or on data management messages. (Salmon et al. 2019)

Blockchains can be used in different settings to provide privacy services. Salman et al. (2019) investigated six approaches that utilize blockchain technology to assess the privacy service. Four of the six approaches used the blockchain platform. Blockchain-based Data Sharing (BBDS) provides privacy for medical records in a cloud environment. It is helpful in lightweight communication systems including user layer, management layer and storage layer. BBDS fulfils the health care records requirement by simplifying the transaction and block header. The system's interactions are secured by identity-based authentication and encryption techniques. This system is implemented in permissioned blockchain. Another proposed approach implemented for a specific IoT use case is the dynamic fair access for IoT. This system lets users register their new resources and define their access policies through a smart contract associated with these resources. A request for a resource made from user A to user B is directed to a blockchain network. Based on the smart contracts resource, the blockchain denies or grants access and sends feedback to the requester who can update their access based on the feedback. Although this approach has been implemented and confirmed for Bitcoin, indicating its feasibility in providing the right access control list management, drawbacks include its lack of real time support, complexity of the blockchain and inflexible implementation.

2.1.1 Loss of control over data due to unauthorized access

Despite the extensive efforts by experts around the globe, issues relating to the security and privacy of the cloud model remain unresolved. Resolving these issues is critical in order to provide viable options to cloud users while gaining the confidence of end users. Unauthorized access to confidential information is the main security and privacy risk in cloud computing. Such unauthorized access can be gained by both internal and external hackers due to the loss of control over the stored data. It may occur with a third-party auditor which is expected to behave in a responsible manner though, in practice, this may not always be the case. (Liu et al. 2017) The loss of control of data is not the only threat arising from unauthorized access, it also impacts control over the data's lifecycle. Data users need assurance that their data is completely removed from the cloud when it is erased - if there a copy of erased data exists, then the act of erasing would be redundant. Such a scenario indicates that data users have no control over the lifecycle of data as more data will be completely erased only when data users change the service provider. (Pearson & Benameur 2010)

2.2 Blockchain as a solution for unauthorized access

One of main feature of blockchain is data anonymity. Blockchain is beginning to address access-control challenges and data confidentiality by providing complete out-of-the-box block data encryption and AAA capabilities. Full encryption of blockchain data ensures that it will not be accessible to unauthorized parties while the data is in transit (especially if data flows through untrusted networks). (Piscini, Dalton, & Kehoe 2017) In blockchain, each user only manages its own keys. Blockchain enhances security by maintaining two distinct ledgers, namely those with individually encrypted data and a transaction ledger which stores encryption access keys to the related data. Each access to the data could be restricted to a certain number of attempts and will be recorded and stored. In other words, if a third party needs to access the data, permission is required to access only specific document at specific times. (Lis and Mendel 2019)

2.2.1 Cross border data transfer

Although cloud computing is a centralized setting, it does not necessarily mean that data are physically stored in a single place. Normally, multiple service provider may participate in delivery of the service (Gai, Raymond, Zhu & 2018) in which the data is transferred between cloud service providers. In fact, such a situation is not ideal since the information that the cloud deals with relates to the personal and sensitive data of users and thus raising concerns over their privacy and security. It also increases the risk of cyber-attack by individuals, organized criminals or governments and opens the door for illegal copying of data as there is no transparency in the cloud regarding the act of transfer. Thus, adequate protection is necessary to protect end users' personal and sensitive data. Different countries have different approaches to protecting the privacy of personal and sensitive data. The new GDPR seeks to resolve the abovementioned issues though problems still persist as illustrated in the following discussion. (Tehrani et al 2018).

2.3 Blockchain as a solution for the issue of data transfer

As explained earlier, blockchains employ a decentralized system as opposed to the cloud's centralized system. Blockchains work by directly transferring data, peer to peer (P2P), without any third party regulating the process. This also means that applicable laws in blockchains will be determined based on the stipulations of the contract since the blockchain is located on a decentralized ledger in a network of many computers instead of in a single server. In terms of ambiguity of the place of formation of the smart contract, courts will need to determine a

method for locating the smart contract's place of formation. One might say that in case of malfunction or unpredictable results, no party can be sued owing to the distributed ledger across a large network, and that no central authority can be held accountable for any failure of the agreement. However, the operator of the blockchain could be held accountable if the malfunction is traceable to their action and in determining jurisdiction and choice of law.

2.3.1 Contractual issues

To ensure compliance of users in cloud services, service providers usually require user candidates to "sign" or consent to an agreement prepared by the provider called a clickwrap agreement. A clickwrap agreement is a contract that appears upon or before accessing a certain computer software or online purchase. It is meant to be a legally binding contract between the two parties and is usually signed by checking an "I agree" box or something similar. (Johansson 2014) The agreement contains information of anything that is important for the user to understand and with terms of use of the application or online purchase. As explained in the following section, while the service provider endeavours to comply with GDPR, (Mackie 2020) it does not mean that there are no issues in the process. Clickwrap agreements are more often seen as a way for service providers to comply with the regulation. However, it lacks any actual "agreement" value on the service user as they compel users to accept the contents of the agreement or not use the service. Users often have a minimal understanding of the terms of the agreement. In most cases, they will just agree on the click wrap agreement without actually reading the whole agreement, making them liable for risks that they are not fully aware of.

2.4 Smart contracts in blockchains

The recent innovation of blockchain technology has gone beyond digital currency applications. One of the most notable functions in the system is the smart contract. A smart contract is a feature that will automatically execute a contract in the blockchain agreed to beforehand, thus eliminating the possibility of deception or fraud after the agreement is concluded. Smart contracts save money and provide autonomy, trust, backup, safety, and accuracy to blockchains. They now provide a mechanism to store, evaluate, and update policies on the blockchain thereby creating new opportunities for people to interact, while reducing the need for intermediaries and costs of transactions. A major element of the blockchain based smart contract is the cryptographic key with which the contract is signed and acknowledged. Parties indicate their consent to the terms of the contract by providing their digital signature utilizing the cryptographic key. This process eliminates intermediaries and strengthens the level of accountability for all parties involved in a way not possible with traditional agreements. Smart contracts have numerous advantages over traditional contracts. As the contracts are translated into computer code, the evidence problem will be lessened since the computer cannot consider external evidence and can only run the code it is given. A smart contract is a powerful form of evidence that will not allow the parties involved to dispute unincorporated issues that are not agreed upon. It overcomes the need to have a person in charge of administrating and executing the contract since its terms can easily be converted to computer code and thereby executed automatically. (Pranata & Tehrani 2022).

3. Conclusion

Despite major attempts to address privacy and security issues, much uncertainty still persists in the cloud environment. The blockchain system, with its emphasis on decentralization, and the collaboration of individual work units for achieving a common goal, can help overcome all the shortcomings of the centralized system. Some of the issues relating to the use of cloud computing services discussed in this article can be resolved by using the blockchain system. The first issue of concern relates to ownership of data and its implications for the possible loss of control over it. This is mainly due to the existence of internal and external attackers and their ability to have unauthorized access to the data. This issue can be resolved by using blockchain which has as one of its main features data anonymity due to its complete encryption in the system.

The second issue is cross-border data transfers since ambiguity regarding the location of data that is stored in the cloud makes it difficult to ascertain the legal jurisdictions that may apply. In the cloud computing system, data - particularly personal data - transferred between multiple service providers is exposed to the risk of being accessed by unauthorized parties. This issue can be eliminated by using blockchains as they provide strict controls and ensure close monitoring over any potential illegal activity.

The third factor relates to contractual issues. Cloud computing usually "forces" people using its services to agree to the terms via a clickwrap agreement. Among the SLA issues is that the service provider may change the term

and conditions, and while some providers may notify the user of the change others may absolve themselves of any such obligation. This can be solved by using the smart contract in blockchain as it is made by consensus of the users, ensuring total understanding between both sides. A clickwrap agreement is non-negotiable and is seen as a take-it or leave-it option. It simply doesn't have the same capacity of blockchain smart contracts in accommodating the needs of the parties involved.

Blockchain is a promising solution for many distributed applications where trust and transparency are critical factors. The Internet of Things (IoT) platform is a multi-layer technology that enables the management and automation of connected devices within its universe. While many research proposals suggest solution for data sharing in IoT platforms by connecting them to blockchain platforms, most follow a hybrid approach where the cloud provider hosts the data and blockchain to ensure trust distribution and integrity and other related issues. These ideas can be used and are still proving the viability of blockchains in addressing privacy and security issues in cloud computing storage. They could be genuine options for data protection as they have the potential to provide robust and strong cyber security solutions where only legitimate users can see the data and where proof of identity is stored in a cryptographic format.

Acknowledgement

This work was supported by scholarship "Stipendien aus Mitteln des ASEA-Uninet, Projektstipendien SP 24" under The Austrian Agency for International Cooperation in Education & Research (OeAD-GmbH).

Reference

- Adrian, A., 2013. How much privacy do clouds provide? An Australian perspective. *Computer Law & Security Review*. 29 (1): 48–57.
- Akbar, A., 2018. Exclusion Clauses and the Reasonableness Test_ what you need to know about the latest court of appeal decisions. <https://gowlingwlg.com/en/insights-resources/articles/2018/exclusion-clauses-and-the-reasonableness-test/> accessed on 15th March 2018
- Banu, S., & Kumar, R., 2018. Security in Cloud Computing Using Blockchain Technology. *AIJR Proceedings*, 1 (NCICCND): 422–439. One of the main features of blockchain is that it eliminates third party interferences between the data inside of the blockchain.
- Bonhomme, V., et al., 2020. Chainlink and iExec collaborate to address the complex off-chain needs of next-generation decentralized applications, <https://medium.com/iex-ec/chainlink-and-iexec-collaborate-to-address-the-complex-off-chain-needs-of-next-generation-702e55ab1ead>
- Cohn, A., et al., 2017. Smart after All: Blockchain, Smart Contracts, Parametric Insurance, and Smart Energy Grids, *George Town Technology Review*.
- Ferdous, M.S., et al., 2017. Decentralised runtime monitoring for access control systems in cloud federations. in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Atlanta, GA, USA, 2632–2633.
- Gai, K., Raymond, K.K., & Zhu, L., 2018. Blockchain-Enabled Reengineering of Cloud Datacenters. *IEEE Cloud Computing*: 21–25.
- Höbl, M., 2011. Cloud Computing: Security and Privacy Issues. https://doi.org/10.1007/978-3-030-11196-0_55
- Hüllmann, T. A., 2018. Are Decentralized Autonomous Organizations the Future of Corporate Governance (Otto Beisheim School of Management). Thesis for: Bachelor of Science in International Business Administration <https://doi.org/10.13140/RG.2.2.34344.88327>
- Johansson, A., 2014. The Enforceability of Clickwrap Agreements. UMEA: UMEA University.
- Lis, P., Mendel, J., 2019. Cyberattacks on critical infrastructure: An economic perspective, *Economics and Business Review* 5 (19): 24–47.
- Liu, B., et al., 2017. Blockchain based Data Integrity Service Framework for IoT data, 2017 IEEE 24th International Conference on Web Services, 25–30 June 2017. 468–475. DOI: 10.1109/ICWS.
- Liu, B., et al. Blockchain based Data Integrity Service Framework for IoT data, 2017 IEEE 24th International Conference on Web Services, 25–30 June 2017. 468–475. DOI: 10.1109/ICWS.2017.54.
- Mackie, J., 2020. Clickwrap in the EU". <https://www.termsfeed.com/blog/clickwrap>
- Madrid Resolution, International Standards on the Protection of Personal Data and Privacy, International Conference of Data Protection and Privacy Commissioners. 5 November 2009. www.privacyconference2009.org/dpas_space/space_reserved/documentos_adaptados/common/2009_Madrid/estandares_resolucion_madrid_en.pdf accessed 6th February 2017
- Microsoft Corp. v. United States* 829 F.3d 197 (2d Cir. 2016)
- Maroufi, M., et al., 2019. On the Convergence of Blockchain and Internet of Things (IoT) Technologies, *Journal of Strategic Innovation and Sustainability* (14) 1. DOI: <https://doi.org/10.33423/jsis.v14i1.990>
- Mungwe, R.E.M., 2017. Legal and Privacy Issues with Cloud Computing in Small and Medium Size Enterprises. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)* 37(1): 44–51.

- OECD., 2014. Cloud Computing: The Concept, Impacts and the Role of Government Policy & quot; OECD Digital Economy Papers, No. 240, OECD Publishing, Paris. <http://dx.doi.org/10.1787/5jxzf4lcc7f5-en>
- Pranata, A. R., & Tehrani, P. M., 2022. The Legality of Smart Contracts in a Decentralized Autonomous Organization (DAO). In *Regulatory Aspects of Artificial Intelligence on Blockchain* (pp. 112-131). IGI Global.
- Pearson, S. & Benameur, A., 2010. Privacy, Security and Trust Issues Arising from Cloud Computing. 2nd IEEE International Conference on Cloud Computing Technology and Science. 693-792.
- Piscini, E., Dalton, D., & Kehoe, L., 2017. Blockchain and Cyber Security. Deloitte.
- Privacy in Cloud Computing. ITU-T Technology Watch Report. March 2012. https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000160001PDFE.pdf.
- Rohrmann, C.A., et al., 2015. Some Legal Aspects of Cloud Computing Contracts, *Journal of International Commercial Law and Technology* (10) 1.
- Salah, K., Rehman, M.H.U., Nizamuddin, N. and Al-Fuqaha, A., 2019. Blockchain for AI: Review and open research challenges. *IEEE Access* 7: 10127-10149. DOI: 10.1109/ACCESS.2018.2890507
- Salman, T., et al., 2019. Security Services Using Blockchains: A State-of-the-Art Survey, *IEEE Communications Surveys & Tutorials* (21) 1.
- Salmon, J., et al., 2019. Fresh Ideas about Business in Emerging Markets. www.ifc.org/thoughtleadership
- Stanciu, A., 2017. Blockchain based distributed control system for edge computing. In 2017 21st International Conference on Control Systems and Computer Science (CSCS) (pp. 667-671). IEEE. DOI: 10.1109/CSCS.2017.102.
- TagElsir, T., et al., 2015. Internal & external attacks in cloud computing environment from confidentiality, integrity and availability points of view. *IOSR Journal of Computer Engineering* (IOSR-JCE) 17 (2): 93-96. <http://www.iosrjournals.org/iosr-jce/papers/Vol17-issue2/Version-5/N017259396.pdf> accessed 6th March 2019.
- Tehrani, P. M., Sabaruddin, J. S. B. H., & Ramanathan, D. A., 2018. Cross border data transfer: Complexity of adequate protection and its exceptions. *Computer law & security review*, 34(3), 582-594.
- Tehrani, P. M., Sabaruddin, J. S. B. H., & Ramanathan, D. A., 2017. The problem of binary distinction in cloud computing and the necessity for a different approach: Positions of the European Union and Canada. *Computer law & security review*, 33(5), 672-684.
- Truong, H., et al., 2019. Towards Secure and Decentralized Sharing of IoT Data, *Proceedings of the 2nd IEEE International Conference on Blockchain*, Atlanta, USA, 2019.
- Zhang, Y., Deng, R., Liu, X. and Zheng, D., 2018. Outsourcing service fair payment based on blockchain and its applications in cloud computing. *IEEE Transactions on Services Computing*.

Context-Aware Cyber Threat Intelligence Exchange Platform

Michael Motlhabi, Phumeza Pantsi, Bokang Mangoale, Rofhiwa Netshiya and Samson Chishiri

Council for Scientific and Industrial Research, Pretoria, South Africa

mmotlhabi@csir.co.za

ppantsi@csir.co.za

bmangoale@csir.co.za

rnetshiya@csir.co.za

schishiri@csir.co.za

Abstract: The ubiquity of network and internet-connected devices has increased exponentially in the past decade. The proliferation of end-user devices has created a lucrative environment for cybercriminals to exploit unsuspecting users at a personal and organizational level. Moreover, businesses and governments are heavily reliant on cyberspace to conduct their business. According to Accenture, in 2019 South Africa saw a spike in cyberattacks on all fronts—banks, Internet Service Providers (ISPs), utilities and eCommerce platforms. This shows that threat actors are continuously looking to exploit new and old vulnerabilities at ever-increasing rates. Furthermore, threat actors are sharing tactics, tools, and procedures to expand their attack surface and to improve the effectiveness of their attacks. Security research tends to be an insular process and rarely do individuals or groups share threat data. This is due to lack of trust, organizational policies, or simply the inability to get the information out to the masses. The idea behind this paper is to design a context-aware threat intelligence exchange platform that encourages collaboration and creates a federated environment amongst different industry stakeholders to share Indicators of Compromise. This paper further aims to define the process of transforming raw Indicators of Compromise into cyber threat intelligence. The platform described in this paper, when implemented, would provide the basic building blocks for developing a highly effective cybersecurity intelligence-sharing system that can improve vulnerability detection and remediation by speeding up the time required to identify/resolve incidents.

Keywords: Security Event Management; Security Information Management, Threat Intelligence, Cybersecurity, Collaboration, Data exchange, Indicators of Compromise, TAXI/STIXX, Tactics, Techniques and Procedures

1. Introduction

There has been significant improvements to technology and the internet over the years, yielding to the enlargement of communication options thereby an increased reliance on Information Technology (IT). In as much as these improvements have changed human livelihood for the better, they have in turn increased opportunities for cybercriminals. South Africa (SA) is currently ranked 34th out of 108 countries, with a score of 0.417 on the global scale for the Cybersecurity Exposure Index (CEI) published in October 2020 (Frisby, 2020). The CEI is based on data collected from the dark web and the deep web's publicly available sources (Cyber Intelligence House, 2020). According to a report by Accenture, SA has seen an increase in cyber-crimes in recent years with notable spikes in 2019 (Accenture, 2020). The report indicates that several sectors in SA were affected by coordinated cyber-attacks from international threat actors who reused some of their attacks. The force of the attacks included stakeholders in banking, e-commerce, telecommunication provides, Non-Governmental Organizations (NGOs) and provincial utility platforms such as water and sanitation. The Experian Security Breach was one of the cyberattacks which occurred in 2020, affecting approximately 24 million and 793000 thousand local businesses (Cimpanu, 2020) (Burger-Smidt, 2020). In 2021, a most recent cyber-attack that occurred in July, South Africa's largest logistics company, Transnet was targeted. The company was hit by a Death Kitty ransomware which disrupted some of its services (CybercrimeSA, 2014-2021) (TechCentral, 2021).

Consequences of cyberattacks do not only include compromise of intellectual property and loss of data but also disruptions to business operations that may lead to financial loss, reputational damage and regulatory fines. According to a financial stability report by the SA Reserve Bank, cyberattacks continually present a danger to the availability of financial services and the functioning of its infrastructures, thereby creating reputational and direct financial risks (South African Reserve Bank, 2021). The increased adoption of Internet of Things (IoT) technology-based processes within the SA government and private sectors has expanded the threat surface as IoT is not secure by default (Allen, 2021). The Operational Technology (OT), industrial control systems (ICS) and supervisory control and data acquisition systems (SCADA) environments are also part of the idea for connectivity and are by design vulnerable to cybersecurity attacks due to the use of legacy IT systems. The most vulnerable industries to cyberattacks in 2021 as reported by CDNetworks, include but are not limited to government

agencies, energy companies, healthcare institutions and small businesses (CDNETWORKS, 2021). Compromise of government agencies could lead to exposure of confidential data such as Personally Identifiable Information (PII) i.e., medical health, military or financial data. SA is heavily dependent on a single entity for electricity supply, thus the higher the risk of impact if that entity is compromised (Allen, 2021).

An effective manner to address cybersecurity threats involves sharing best practices, adopting codes of practice for threats and the exchange of information throughout the various sectors. The more that is known about a threat, the easier it is to understand, track and counter it (Lauren & Irwin, 2015). This paper details a Proof of Concept (PoC) for a cyber threat sharing platform that uses best practices and protocols for threat data exchange and definitions in the SA cyber context. Chapter 2 expands on the current state of the art in the Cyber Threat Intelligence (CTI) domain. Chapter 3 discusses the threat exchange methodology for the PoC. Chapter 4 explains the potential framework for a context-aware CTI system. Chapter 5 focuses on standards and legislative compliance in the SA context and Chapter 6 concludes the paper and provides possible future work and improvements.

2. State of the Art for Threat Intelligence Platforms

(Kaspersky, 2021) defined threat intelligence as a method of filtering through large amounts of data to identify and analyze cyber threats that can potentially cause harm. A threat intelligence platform aims to provide organized, analyzed, and refined information regarding potential and current risks that threaten various ICT entities within defined cyberspace (Cascavilla, et al., 2021). The intelligence from these platforms can be embedded into organisational workflows and assist decision-makers in making tactical decisions. Some examples of threat intelligence platforms include:

- **The Malware Information Sharing Platform (MISP)** – Open-source threat intelligence platform for gathering, sharing, storing, and correlating Indicators of Compromise (IoC) of targeted attacks, threat intelligence, financial fraud information, vulnerability information or even counter-terrorism information (MISP Project, 2021).
- **The Collective Intelligence Framework (CIF)** – Actionable threat intelligence platform focused on speed, performance, and integration (GADGETS, 2021).
- **ThreatQ** – Threat intelligence platform aimed at providing defenders with the context, customization and collaboration needed for increased security effectiveness and efficient threat operations and management (ThreatQuotient, 2021)
- **Anomali ThreatStream** - Threat Intelligence Management that automates the collection and processing of raw data, filters out the noise and transforms it into relevant, actionable threat intelligence for security teams (Anomali, 2021)

Cyber threat intelligence sharing provides a proactive manner of creating situational awareness through information sharing, regarding new threats and vulnerabilities. However, this exchange of information heavily relies on trust between stakeholders. A literature survey conducted by (Wagner, et al., 2019), discusses various factors that may hinder or support the process of threat intelligence sharing between stakeholders. These include automation, legal factors, policies, cultural differences, as well as the human role in threat intelligence sharing.

A report published by (ENISA, 2018), presented some of the limitations related to the state and usage of threat intelligence platforms. The authors stated that the amount of data shared tends to be large because of combining data from open-sources, commercial sources as well as private communities. In addition to a large amount of data, there is limited technology enablement to allow end-users to efficiently facilitate the relevancy determination process. Moreover, these platforms have limited analysis and management capabilities to allow users to interpret, analyse, enrich, and react to the threat information received promptly.

The threat intelligence platform analysis done by (Alessandra, et al., 2020), revealed that most of the threat intelligence platforms focus on various specific features such as speed, performance, and visualization of information. These features are dependent on the goal of the platform and as diverse as they are, they are all considered to be threat intelligence platforms mainly because there is no standardized definition for the concept of cyber threat intelligence.

3. Cyber threat information exchange

This section describes the architecture of the repository used to exchange Cyber-Threat Intelligence (CTI). CTI is a collection of intelligence combining Open-Source Intelligence (OSINT), Social Media Intelligence (SOCMINT), Measurement and Signature Intelligence (MASINT), Human Intelligence (HUMINT), and Technical Intelligence (TECHINT) from deep and dark Web (Sari, 2018). Various literature works have demonstrated that CTI can potentially reduce cybersecurity risks and strengthen cyber resilience in both simple and complex cyberspaces (Mtsweni, et al., 2016). Therefore, for any CTI platform to be adequately effective, information sharing about cyber threats must be conducted within a federated community and following clear standards to facilitate automation.

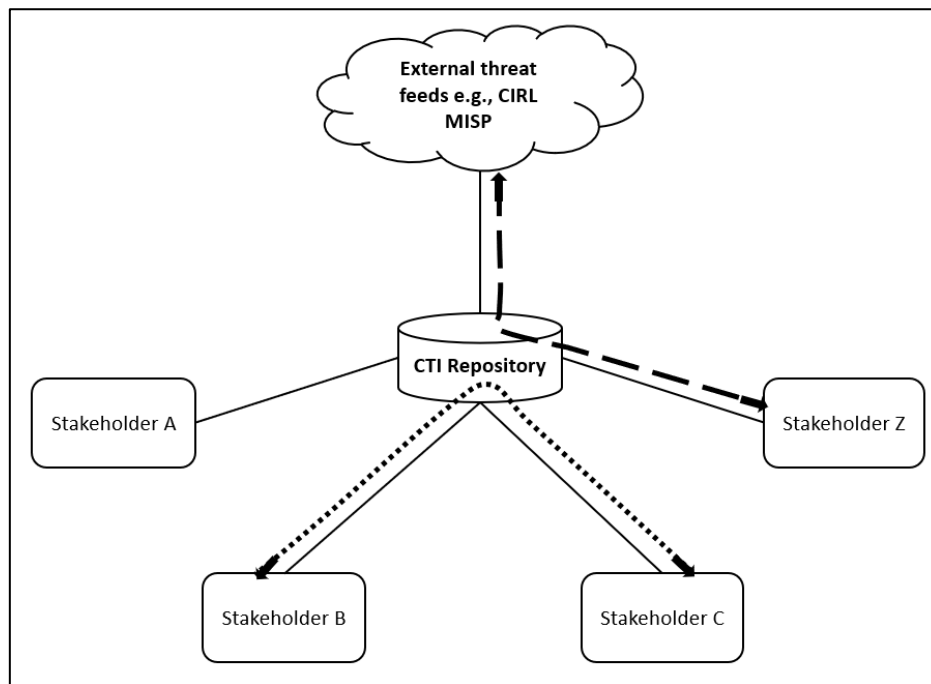


Figure 1: Distributed hub and spoke architecture implementation

Figure 1 shows an architecture that depicts the potential information exchange process of security-related data. Additionally, Figure 1 indicates that a CTI exchange platform should implement a hub and spoke architecture. Hub and spoke is a sharing model in which there is a main data hub (CTI Repo as per Figure 1) where information is coordinated and shared with CTI Stakeholders as seen in Figure 1. The hub and spoke architecture is chosen for the web application because it is advantageous to have a central place for gathering data, and it allows for scaling (Rudman & Irwin, 2016). The exchange of security incidents in this methodology is bi-directional, with entities alternating between producer and consumer CTI stakeholders. The entity publishing incidents to the central CTI Repo at a particular instance, is the producer of CTI and the one subscribing/receiving from the CTI Repo, is the consumer CTI.

In this model, CTI will be exchanged using the Trusted Automatic Intelligent Information Exchange (TAXII) protocol (OASIS Cyber Threat Intelligence (CTI) TC, 2021). The TAXII protocol will deliver cyber threat intelligence in a simple and scalable manner. Since the CTI exchange platform uses TAXII, stakeholders can choose what information they share and with whom they share CTI with, thus increasing the flexibility of the platform. TAXII supports different sharing models like the hub and spoke, which should be implemented in the CTI exchange platform. TAXII defines the “how”, in this implementation, the Hypertext Transfer Protocol Secure (HTTPS) protocol is utilised for transportation because of its security properties.

There are currently no standard formats for IoCs defined, however, over the past decade, security experts have developed their formats, such as Incident Object Description Exchange Format (IODEF), OpenIOC3, Cyber Observable Expression (CybOX) and Structured Threat Information eXpression (STIX) (Rudman & Irwin, 2016). STIX is used to describe information about cyber threats. The goal of STIX is to have a language or representation that allows information to be easily stored, analysed and shared in a consistent manner amongst stakeholders.

STIX not only allows for the creation of indicators but also other contextual information such as observables, incidents, exploit targets, courses of action, campaigns and threat actors (The Mitre Corporation, 2016). IoCs are data that is found in system log entries or files that identify potentially malicious activity in cyberspace (Lord, 2020). The following IoCs will be used to detect threats:

1. IP addresses - geographic irregularities, such as traffic from countries or locations where the organization does not have a presence.
2. High authentication failures – attackers use automation to authenticate using phished credentials. A high rate of failed authentication attempts could indicate that an attacker is using multiple incorrect credentials to find credentials that will provide access to a network.
3. Flooded traffic to a specific site or IP – a compromise on a network could make devices create or form part of a botnet, to deploy Distributed Denial of Service (DDoS) attacks.
4. Unusual activity from administrator or privileged accounts, including requests for additional permissions.
5. Domain Name System (DNS) requests anomalies - DNS requests and traffic to command-and-control servers often follows a standard pattern which can serve as a good indicator for suspicious activity.
6. Common Vulnerability and Exposure (CVE) number - CVE number provides a standardized identifier for a given vulnerability or exposure. These can be used to identify vulnerabilities that could be exploited.

The STIX and TAXII protocols provide the CTI exchange platform to leverage the definitions of what are relevant IoCs and how they can be shared amongst different stakeholders in the federation. The information being shared by the CTI exchange platform will be in the STIX format and displayed in Hypertext Transfer Markup Language (HTML).

In a web application's Graphical User Interface (GUI), users will gain access to the CTI exchange platform, where they can classify which stakeholder group they are federated to during the sign-up process. Once users have a profile, they can follow relevant organizations and get notified if those organizations upload new threat information. The user profile will determine what CTI will be reported to the user, based on three classes labelled below:

- Strategic – for visualizing broader trends for non-technical audiences.
- Tactical – to outline tactics, techniques and procedures of threat actors and must include technical context about remediation protocols.
- Operational – includes technical details about specific attacks and information about what attack vectors are being used.

To facilitate information exchange between stakeholders in different organizations, MISP will be used (NCI Agency, 2017). The MISP platform is used for sharing IoCs of targeted attacks for only NATO member nations, governmental and cyber defence organisations. MISP allows for the uploading of information (via text or template), to be manually completed or the PyMISP/ZeroMQ can be used for automation. MISP allows for data export by generating Intrusion Detection System (IDS) rules like Snort and IoC formats such as STIX. The platform discussed in this paper is different because it focuses on two key aspects labelled below:

- Southern African Development Community (SADC) – CTI will be shared with stakeholders in the SADC region. This is to increase the cyber security posture of this region. (Accenture, 2020) reports that there has been a visible increase in cyber-attacks in the Southern African region.
- Context-aware network-level information is collected and compared with stakeholders to provide the most relevant and accurate threat response and reporting.

4. Context-Aware cyber-threat intelligence

A context-aware cyber-threat intelligence system has visibility to the network elements that are present in the internal environment. Its objectives are to facilitate information sharing, enable automation, and facilitate the generation, refinement and vetting of data through burden-sharing collaboration. A context-aware threat intelligence platform can support timely decision-making, and automate responses to the greatest extent possible, to limit the impact of a potential threat. With this methodology, security teams can get the information they need quickly and from a single source in a rich and contextualized format.

4.1 High-level system overview

Network environments usually generate large amounts of security logs from different network devices and applications. These logs can be collected from internal devices, and cyber-related information can be collectively mapped into cyber intelligence. The platform proposed in this paper, will also consume data from external cyber threat feeds such as MISP, Malware, CVE, malware sites etc. The system will then ingest both the internal and external cyber information to provide a security posture, that is contextualized to a specific organization. The organizations' security posture is shared with key stakeholders who might be impacted. During the CTI sharing process, if any logs are shared between stakeholders, the shared logs will be anonymized.

Since the platform collects network data, the design must allow a stakeholder organization to store cybersecurity data that can either be kept private or shared with others. When data is identified as being private, the system should ensure that the data is never made publicly available to comply with privacy laws and regulations.

4.2 Navigation and flow

Figure 2 shows several use cases that the CTI exchange platform can execute. The web application first prompts the user to sign up, and log in, navigate to the dashboard page where the user can perform six major actions. The six capabilities are, Notification and Alarming, Reporting, Collections, Risk Management, Search and Topology. The Administration capability is only accessible from the central CTI Repo (see Figure 1) local instance and is primarily used to configure the platform, check the system health status, and manage licensing.

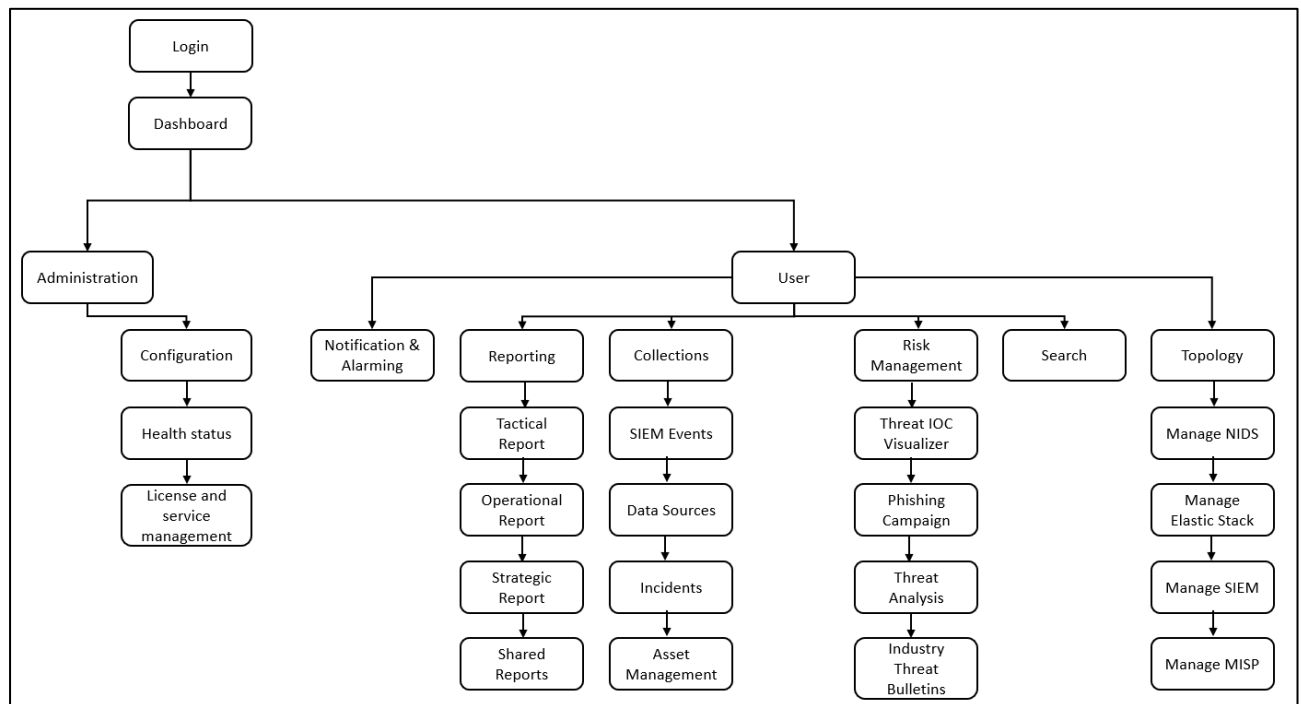


Figure 2: CTI web application capability flow diagram

4.3 Functional architecture

The main function of the threat intelligence system discussed here is to create an eco-system that is capable of efficiently capturing IoCs and other indicators. A central database will allow for the storage of technical and non-technical information about malware samples, incidents, attackers, and intelligence. The development of a CTI exchange platform is divided into three categories described in Figure 3.

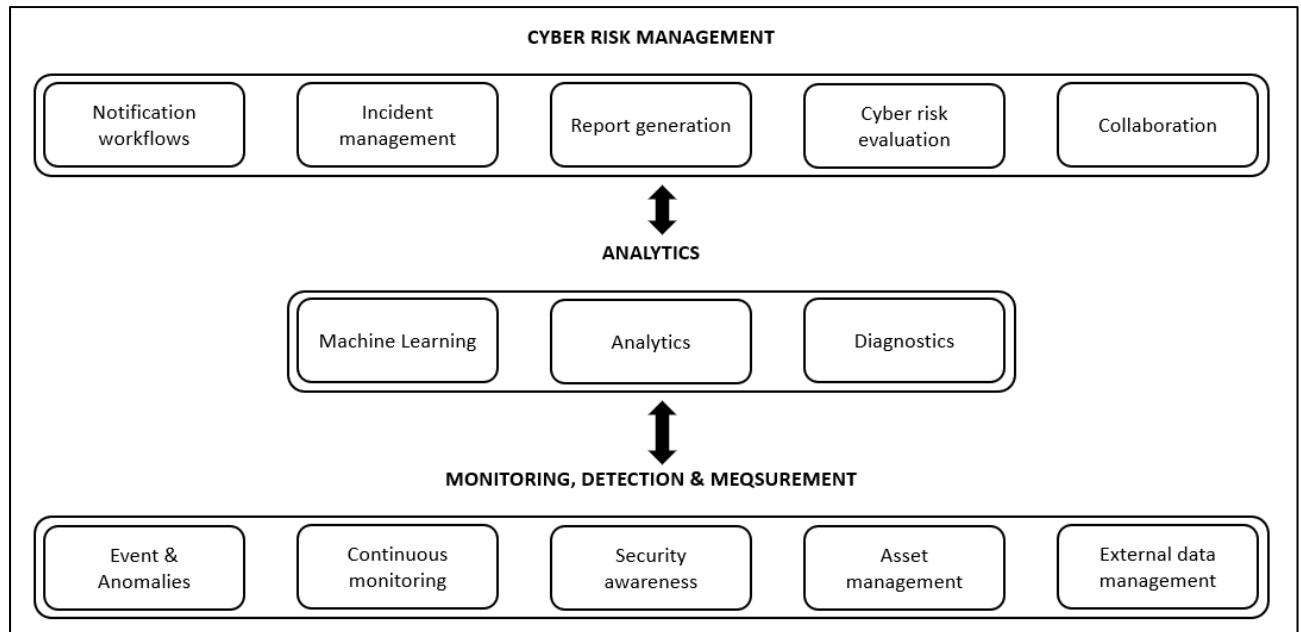


Figure 3: Functional architecture for a context-aware CTI exchange platform

Monitoring, Detection and Measurement: In this segment of the platform, local network variables are collected that are unique to this environment. This enables the CTI sharing platform to accurately monitor attributes that are relevant to the local environment for analysis. Attributes such as network elements events and anomalies form part of the critical system dependencies, that are used to evaluate the security posture in the analysis phase.

Cyber Risk Management: under this category, five distinct attributes can inform the system user about the cybersecurity state, these are functionally explained as follows:

- **Cyber Risk Evaluation** – a score is generated (in line with CVE scores that measures the likelihood of vulnerabilities being exploited).
- **Notification workflows** – the detection of a high severity IoC will action a notification that will be shared with the stakeholders detailing the specifics of the threats, actors, tactics and/or techniques (Dandurand & Serrano, 2013).
- **Report Generation** – a report detailing the threat and the means to remediate that threat will be compiled into three types of reporting for strategic, tactical, and operational audiences.
- **Collaboration** – different stakeholders will be able to share all the threat intelligence information automatically through the client application installed in the network. The exchanges are necessary to achieve common goals in federated environments and to exploit collaboration opportunities. This objective of quick collaboration and action to remediation is not a technical issue and thus is beyond the scope of these works. It is assumed in this paper that all the stakeholders will comply with the protocols to respond and share their environment specific threat intelligence as quickly as they can respond to them.

Analysis: as data is collected from both the internal environment and the external malware platforms such as the MISP. Machine Learning is used for data analysis and transformation into threat intelligence to find network anomalies (Motlhabi, et al., 2021). IoCs are then extracted to show what pertinent threats exist that could compromise the security of the network and other external stakeholders. This section also performs diagnosis, where remediation protocols are defined, reports are populated, and potentially affected stakeholders are notified of the cyber threats.

4.4 System SIEM process

A context-aware threat intelligence system should have the capability to analyse local network data. This local analysis ability provides a mechanism for integrating external source threat feed intelligence to provide a security posture that is relevant to the organisation it is deployed in. As such a Security Event Management (SEM) can perform threat monitoring, event correlation, and incident response by analysing the log and event

data in real-time. Security Information Management (SIM) performs collection, analysis, and reporting on log data (Software Testing Help, 2021). Figure 4 shows a logical set of steps that the platform will follow during its Security, Information and Event Management (SIEM) phase.

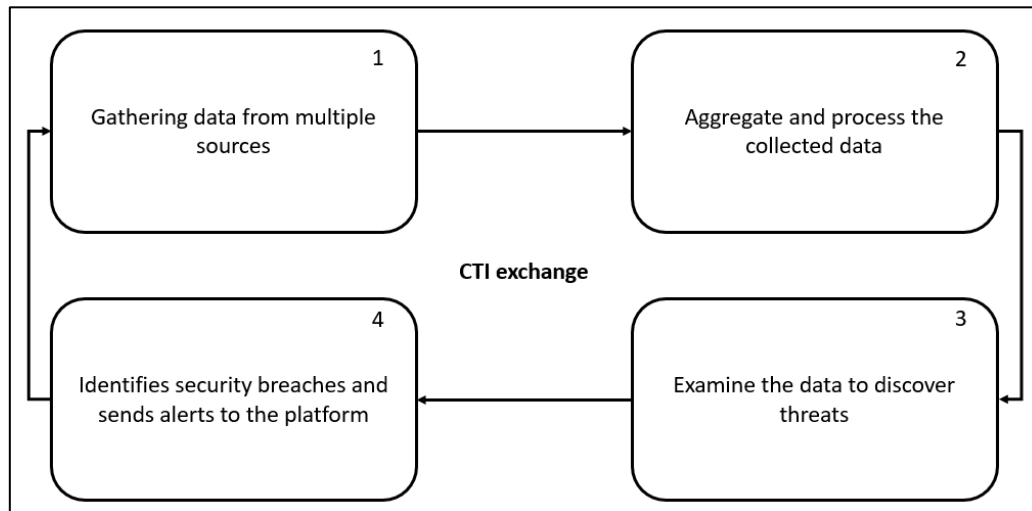


Figure 4: Need to redraw this picture to include external sources

Step 1 cyber threat intelligence data such as malware information, phishing data, IoC and other indicators are collected from two sources i.e., the organisations' internal network and the external cyber threat sources that have been collected by the system and/or shared by other organisations. Storage of cybersecurity indicators in a structured formal way is done in this step.

Step 2 raw and rich data is brought together from different sources as public Open-Source Intelligence (OSNIT) feeds and form part of the default configuration of the SIEM function. The threat feeds are then used as a source of correlations for all the events and attributes without the need for the organization to import them directly into their system. Cyber threat feeds will be collected as Comma Separated Values (CSV), free text or JSON. CSV will allow for the specific cyber attributes to be selected before processing while free-text formats provide unstructured indicators to be added to the threat intelligence system. At this level, advanced filtering capability will be enabled and applied by end-users.

Step 3 perform an analysis using supervised ML methods for the identification and categorization of incidents and events.

Step 4 is where security breaches are identified with the aid of intuitive GUIs for end-users to collaborate on events and threat indicators. The GUI helps navigate seamlessly between events and their correlations but also shows alerts and remediation suggestions. In step 4 the technology shows security orchestration and automation where threats can be automatically blocked that either the stakeholders or the community have blocked.

Figure 4 showed how threat data is collected, processed, analysed and shared with other organizations. Threat data is automatically made available to participating organisations via the client GUI that is installed at the organisation's premises. The exchange is made possible by threat sharing protocol STIX which defines what intelligence is shared (Sauerwein, et al., 2017). The platform can thus synchronize threat intelligence and distribute it to different application clients seamlessly. Additionally, the platform will share its data with other threat repositories such as the MISP or the CIRCL. Since the data collected is structured, many existing incident management and SIEM tools can ingest this data directly.

5. Standards and Legislative Compliance

Various cybersecurity standards and frameworks address security hardening, incident response etc. For example, the ISO 27001 standard provides specifications that assist organizations in designing, implementing, reviewing, maintaining, and continuously improving an ISO 27001 compliant ISMS (Bawono, et al., 2020). The National Institute for Standards and Technology (NIST) cybersecurity framework provides broad guidelines and

best practices that ensure a well-built, and continuous improvement of any organization's security posture (Gordon, et al., 2020). Security activities within the NIST framework are categorised according to five basic functions. These are Identify, Protect, Detect, Respond, and Recover (see Table 1). The CTI exchange platform discussed in this paper is designed to support these initiatives and form collaboration between collaborators by sharing IoCs, Tactics, Techniques and Procedures (TTP), remediations etc.

Table 1: NIST Framework Functions

Function	Description
Identify	The Identify focuses on identifying critical organization assets and processes.
Protect	The protect function involves specifying, as well as implementing security mechanisms to protect the critical assets and processes.
Detect	The organizations detect function outline activities to identify occurrences of cyber-attacks.
Respond	The organization's response function focuses on procedures to be followed once a cyber-incident is detected, this includes but is not limited to roles and responsibilities and limiting the impact of the cyber incident.
Recovery	The organization's recovery function conducts recovery planning processes and procedures to restore systems and/or assets damaged by cybersecurity events.

In 2015 the SA government published the National Cybersecurity Policy Framework (NCPF). Its purpose was to establish a safe and secure cyber environment that emphasizes the protection of critical information infrastructure. For this to be achieved a joined force between different stakeholders (civil society, private, and public sectors respectively) is necessary, in understanding as well as addressing risks within cyberspace. (Fonseca, et al., 2021) Through the guidance of NCPF, a national cybersecurity hub, which is SA's National Computer Security Incident Response Team (CSIRT) was established. The CSIRT identifies and counters cyber threats.

SA's Protection of Personal Information (POPIA) is strictly intended to set out conditions on how and when an organization is allowed to process ("including collecting, using, transferring, matching and storing" (Rachel, et al., 2021)) personal information. SA organizations are required to ensure that they comply with the ACT by July 2021. The ACT consist of three role players, that is the Data Subject, Responsible Party, and the Operator. Amongst its compliance requirements, the POPIA specifies that responsible parties should report data breaches, only share information when they are lawfully allowed, and verify if they are lawfully authorized to transfer PII. The technology presented here will not be consuming any PII from the networks it is deployed on. Furthermore, the organization decides what information it will share with other collaborators.

6. Conclusion

Cybercriminals are sharing tools and information. Cybercriminals also sell exploits and tools to each other. Ransomware-as-a-service. Organizations in the same industries with common goals should do the same to defend against cyber-attacks, the platform presented here shows that such a system could be developed. Organizations will have a way to share cybersecurity intelligence, while their private data remains private, and in some cases should an organization wish to share their data they can decide what to share and with whom. This intelligence will provide a contextualized security posture about the network environment and specific threats that are relevant to each industry. The collaboration amongst the industry players would yield higher defences against threat actors. Since their TTP's would have been shared and exposed to a wider audience. Our system will have the following capabilities: store, analyse, share, collaborate on cybersecurity indicators, malware analysis, and use the IoCs and information to detect and prevent attacks or threats against ICT infrastructures and organisations. While providing stakeholder and expert level reporting and analysis. A CTI exchange platform should reduce data duplication and the effort required for the analysis and classification of potential threats. The solution's main function is to create a CTI repository in SA which will assist collaborators with identifying, protecting, and responding to cyber threats related to their industry. Future work will involve the implementation of how the gathered intelligence can be converted to different formats that can be used in a Security Operations Centre (SOC) together with various threat detection and analysis tools. Improving the GUI to cater for specific industries with similar interests. To continuously find better ways to improve threat detection and intelligence sharing with stakeholders.

References

Accenture, 2020. *Insight into the Cyberthreat Landscape in South Africa*, Johannesburg: Accenture.

- Alessandra, d. M. e. S., João José Costa, G., Robson, d. O. A. & Villalba, L. J. G., 2020. A Methodology to Evaluate Standards and Platforms. *Future Internet*, Volume 12, p. 108.
- Allen, K., 2021. *Critical infrastructure attacks: why South Africa should worry*. [Online]
Available at: <https://issafrica.org/iss-today/critical-infrastructure-attacks-why-south-africa-should-worry>
[Accessed 30 07 2021].
- Anomali, 2021. *Anomali threatstream*. [Online]
Available at: <https://www.anomali.com/products/threatstream>
[Accessed 25 May 2021].
- Bawono, A. et al., 2020. Analysis Corellation of the Implementation Framework COBIT 5, ITIL V3 and ISO 27001 for ISO 10002 Customer Satisfaction. In: s.l.:EasyChair.
- Burger-Smidt, A., 2020. *Massive data breach in SA – did Experian do enough?*. [Online]
Available at: <https://www.bbrief.co.za/2020/09/09/massive-data-breach-in-sa-did-experian-do-enough/>
[Accessed 25 July 2021].
- Cascavilla, G., Tamburri, D. A. & Van Den Heuvel, W.-J., 2021. Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, Volume 105, p. 102258.
- CDNETWORKS, 2021. *THE INDUSTRIES MOST VULNERABLE TO CYBER ATTACKS IN 2021*. [Online]
Available at: <https://www.cdnetworks.com/cloud-security-blog/the-5-industries-most-vulnerable-to-cyber-attacks/>
[Accessed 30 07 2021].
- Cimpanu, C., 2020. *Experian South Africa discloses data breach impacting 24 million customers*. [Online]
Available at: zdnet.com/article/experian-south-africa-discloses-data-breach-impacting-24-million-customers/
[Accessed 25 July 2021].
- Cyber Intelligence House, 2020. *Cyber Exposure Index*. [Online]
Available at: <https://cyberexposureindex.com/>
[Accessed 31 07 2021].
- CybercrimeSA, 2014-2021. *Cybercrime.org.za*. [Online]
Available at: <http://cybercrime.org.za/>
[Accessed 28 07 2021].
- Dandurand, L. & Serrano, O. S., 2013. *Towards Improved Cyber Security Information Sharing*, Tallinn: 5th International Conference on Cyber Conflict.
- ENISA, 2018. *Exploring the opportunities and limitations of current Threat Intelligence Platforms*, s.l.: ENISA.
- Fonseca, R., Wyk, S. v. & Jo-Ansie, 2021. Cybersecurity in South Africa: Status, governance, and prospects. In: *Routledge Companion to Global Cyber-Security Strategy*. s.l.:Routledge.
- Frisby, J., 2020. *Cybersecurity Exposure Index (CEI) 2020*. [Online]
Available at: <https://passwordmanagers.co/cybersecurity-exposure-index/>
[Accessed 31 July 2021].
- GADGETS, C., 2021. *The FASTEST Way to Consume Threat Intelligence..* [Online]
Available at: <https://csirtgadgets.com/collective-intelligence-framework>
[Accessed 4 August 2021].
- Gordon, A. L., Loeb, M. & Zhou, L., 2020. Integrating cost--benefit analysis into the NIST Cybersecurity Framework via the Gordon--Loeb Model. *Journal of Cybersecurity*.
- Kaspersky, 2021. *Threat Intelligence Definition. Why Threat Intelligence Is Important for Your Business and How to Evaluate a Threat Intelligence Program*. [Online]
Available at: <https://www.kaspersky.com/resource-center/definitions/threat-intelligence>
[Accessed 31 July 2021].
- Lauren, R. & Irwin, B., 2015. *Towards a framework for automated generation*, Grahamstown: Rhodes University.
- Lord, N., 2020. *Digital Guardian*. [Online]
Available at: <https://digitalguardian.com/blog/what-are-indicators-compromise>
[Accessed 11 August 2021].
- MISP Project, 2021. *MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing*. [Online]
Available at: <https://www.misp-project.org/>
[Accessed 03 August 2021].
- Motlhabi, M., Panti, P. & Netshiya, R., 2021. *A Machine Learning Deep-dive Analysis into Network logs*, Tennessee, US: ICCWS 2021, 6th International Conference on Cyber Warfare and Security.
- Mtsweni, J., Mutemwa, M. & Mkhonto, N., 2016. Development of a cyber-threat intelligence-sharing model from big data sources. *Journal of Information Warfare*, 15(3), pp. 56-68.
- NCI Agency, 2017. *NCI Agency Malware Information Sharing Platform*, s.l.: NCI.
- OASIS Cyber Threat Intelligence (CTI) TC, 2021. *Introduction to TAXII*. [Online]
Available at: <https://oasis-open.github.io/cti-documentation/taxii/intro.html>
[Accessed 04 08 2021].
- Rachel, A. et al., 2021. POPIA Code of Conduct for Research. *South African Journal of Science*, 117(5-6).
- Rudman, L. & Irwin, B., 2016. *A sharing platform for Indicators of Compromise*. Fancourt, George, Southern Africa Telecommunication Networks and Applications Conference (SATNAC).

- Sari, A., 2018. *Context-Aware Intelligent Systems for Fog Computing Environments for Cyber-Threat Intelligence*. s.l.:Springer, Cham.
- Sauerwein, C., Sillaber, C., Mussmann, A. & Breu, R., 2017. *Threat Intelligence Sharing Platforms: An Exploratory Study of Software Vendors and Research Perspectives*. Leimeister, s.n.
- Software Testing Help, 2021. *Top 11 Best SIEM Tools In 2021 For Real-Time Incident Response And Security*. [Online]
Available at: <https://www.softwaretestinghelp.com/siem-tools/>
- South African Reserve Bank, 2021. *FINANCIAL STABILITY REVIEW*. [Online]
Available at: <https://www.resbank.co.za/en/home/publications/publication-detail-pages/reviews/finstab-review/2021/First-edition-2021-Financial-Stability-Review0>
[Accessed 29 07 2021].
- TechCentral, 2021. *Transnet likely hit by Death Kitty ransomware attack*. [Online]
Available at: <https://techcentral.co.za/transnet-likely-hit-by-death-kitty-ransomware-attack-experts/109642/>
[Accessed 05 08 2021].
- The Mitre Corporation, 2016. *The Mitre Corporation*. [Online]
Available at: <http://stixproject.github.io/usecases/>
[Accessed 19 08 2021].
- ThreatQuotient, 2021. *THREATQ™ Platform*. [Online]
Available at: <https://www.threatq.com/threatq-platform/>
[Accessed 18 June 2021].
- Varonis.com, 2021. *134 Cybersecurity Statistics and Trends for 2021*. [Online]
Available at: <https://www.varonis.com/blog/cybersecurity-statistics/>
[Accessed 19 08 2021].
- Wagner, T. D., Mahbub, K., Palomar, E. & Abdallah, A. E., 2019. Cyber threat intelligence sharing: Survey and research directions. *Computers & Security*, Volume 87, p. 101589.

The Development of Cybersecurity Awareness Measurement Model in the Water Sector

Bryan S. Mufor, Annlizé Marnewick and Suné von Solms

Postgraduate School of Engineering Management, Department of Electrical Engineering Science, University of Johannesburg, South Africa

amarnewick@uj.ac.za

svonsolms@uj.ac.za

Abstract: Cyber-attacks are one of the main threats to information systems, and humans have been identified as the weakest link with regards to information security. This study aims to develop a measurement instrument to evaluate the level of cyber security awareness (CSA) in the water sector in South Africa. There are lots of synergies with regards to cyber system usage across industries, and as a result this study will take a broad base approach in configuring an instrument that can be used to adequately assess the sample space in question. Having a reliable instrument to measure cyber security awareness helps mitigate the failed attempts at preparing employees for imminent cyber disruptions by pin-pointing areas where the training is needed before campaigns can be organised. This study will show that the psychology of employees with respect to cyber security awareness is compartmentalised into three traits: knowledge, attitude, and behaviour. These three traits were assessed under the following eight focus areas to check employee resilience to cyber security: IS policy adherence, Password management, Email use, Internet use, social media use, mobile devices, information handling, and incident reporting. In practice, employees will be required to answer questions formulated under these focus areas to evaluate their cyber security awareness (CSA) level. The model proposed in this paper was developed to test cybersecurity awareness in the water sector, but can be utilised in other sectors for cybersecurity awareness testing.

Keywords: Cybersecurity; Awareness; Critical Infrastructure

1. Introduction

The age of digitalisation has brought with it an improved level of connectivity and automation. Organisations across the globe have taken full advantage of the digitisation in many ways including, but not limited to, improved access to communication, efficient scaling of operations, and overall cost reduction (Hasan *et al.*, 2021). However, this expansion unfortunately increased the risk and level of exposure of users of interconnected web information systems (Rantapelkonen and Salminen, 2013; Strupczewski, 2021). This rapid growth accentuated the need for innovative methods to protect information systems from malware attacks (Rantapelkonen and Salminen, 2013). Information technology (IT) security is not only an imperative tool for modern organisations, but also a means to retain competitive advantage over the competition in industry (Kong, Kim and Kim, 2012), and improving awareness and readiness to handle cyber-attacks across organisations can have positive effects on the performance of organisations in financial, non-financial and security sectors (Hasan *et al.*, 2021).

Cyber security awareness is an important aspect to evaluate by organisations to fortify cyber security protocols and configurations. This study aims to develop a comprehensive instrument to measure cyber security awareness in the water sector using previously studies psychological inclinations of employees.

2. Cyber security considerations for water sector

This study looks in the CSA levels of employees in the water sector because this sector is among the main targets of cyber-attacks (The White House, 2013). Traditional water systems generally operate on the SCADA system, which monitors a variety of processes along the value creation chain, including raw water extraction and/or collection, transport of water, monitoring and control of purification process, treated water distribution, and control of pressure boost pumps (Luijff, 2008).

SCADA systems have largely been dependent on customised operational technology—technology which requires in-person operation (Hassanzadeh *et al.*, 2020). However, according to Hassanzadeh *et al.* (2020), these SCADA infrastructure are increasingly overlapping operational technology with information technology systems, essentially exposing them remote or cyber tampering. SCADA systems that integrate IT into their operations tend to increase vulnerabilities in the system. However, due to the dispersed nature of water systems, employees have little choice but to access these system remotely to perform operational tasks (Hassanzadeh *et al.*, 2020), and if these individuals are compromised, the entire system becomes vulnerable to pending attack.

Third party personnel such as SCADA manufacturers, who are allowed to perform maintenance on SCADA systems, and who are usually given full access to these system are another source of risk. Luijff, Ali and Zielstra (2011) indicate that managers in the water sector have raised this as an area of vulnerabilities. There exists many incidents where a lack of CSA can leave critical infrastructure, including water, vulnerable to cyber attacks. In many of the cases, the areas of vulnerability are simple issues, such as are password management, email use, internet use, and incident reporting.

3. Methodology

The purpose of this study is to develop a comprehensive measuring instrument to determine CSA of employees in the water industry. The study was conducted in four phases. A systematic literature review was conducted on publications by experts in the field to assess the tools that they have developed in the past, and how these tools were applied to study CSA levels (phase 1 and 2). The data from the identified literature was extracted and assessed (phase 3). This information was utilised to develop a measurement model based on other CSA tools previously tested in industry (phase 4). The following steps were followed in the development of the CSA measurement tool:

- Phase 1: Literature wide scope search, which included the initial search for publications from databases, application of exclusion criteria, screening by title and removal of duplicates and restricted content.
- Phase 2: Literature final content selection, which included validation of publication via abstract, full text quality appraisal process and backward referencing to identify additional content.
- Phase 3: Comparison of measurement theories, which included the identification of model base theory, selection of focus areas and measurement traits, as well as the identification of organisation type and status considerations.
- Phase 4: CSA model design, which included the selection of base theory, focus areas and measurement traits. It also included the consideration of water specific aspects and the final model development.

The execution of these phases will be discussed in detail in the following sections.

4. Search strategy and selection of relevant material

To avoid a random and distorted search process, a comprehensive search strategy was employed, to assure the reader of the level of diligence applied to complete the work presented (Meades, 2015). Systematic literature reviews have strict requirements for implementation of search strategies and the selection of relevant material (Snyder, 2019).

4.1 Phase 1: Literature wide scope search

To identify academic material that is considered relevant to the research objective, targeted searches were conducted using key terms often applied in the field of cyber security. The key terms were identified using a variation of the PICO (People, Intervention, Comparators, Outcome) model (Purssell and McCrae, 2020). This variation entails exposure (E) in place of intervention and the removal of Comparators - PEO. The question to be answered was: *What are the current instruments available to measure CSA in industry in general?* In accordance with the PEO model, the following answers were used to determine the search terms, where the final terms are shown in **Table 1**:

- **People** (representing the sample space under investigation): industry.
- **Exposure** (representing the level of exposure of people): cybersecurity awareness.
- **Outcome** (represents the outcome required): instrument to measure awareness levels.

Table 1: Key search terms and synonyms

Population	Exposure	Outcome		
<i>Industry(ies)</i>	<i>Cybersecurity</i>	<i>Awareness</i>	<i>Measurement</i>	<i>Instrument</i>
Organisation(s) Company(s) Employee Worker(s) Human Factor	Information security	Education Culture	Assessment Evaluation	Model Mechanism Tool Framework Questionnaire Study

For this study, six academic databases were explored to find all relevant publications that could assist in the development of a CSA measurement tool, which included Association for Computing Machinery Digital Library (ACMDL), Emerald, IEEE Xplore, ScienceDirect, SpringerLink and Taylor & Francis Online.

The initial search recorded a total of **24596** titles. After every search, a set of exclusion criteria was applied, which included: papers not published in English, papers not peer-reviewed, magazine and newspaper articles, papers not in used subscription lists, and papers published before 2011. After the exclusion of papers not adhering to the abovementioned criteria, a total of **5928** papers were retained. These papers were screened via titles, removing any papers which do not focus on CSA. After the removal of duplicates and papers which could not be accessed, a total of **120** papers were retained for full content selection.

4.2 Phase 2: Literature final content selection

After the completion of Phase 1 of the search process **120** publications were retained. The retained publications were screened by abstract. All publications that did not specifically discuss the development of a measurement instrument for CSA in the abstract were excluded. A total of **25** papers were retained after the abstract validation step.

For the full text analysis, the following research analysis question was used for the full text quality appraisal process: *Did the author develop an instrument to measure CSA levels?* This quality appraisal was used to determine to what degree the selected studies meet the criteria of the current study—the degree to which the studies answer each predefined research analysis questions were scored on a 5-point scale, ranging from irrelevant (1) to inch-perfect (5). A further 7 publications were excluded, leaving a total of **18** publications. Backward referencing was utilised to identify relevant content. Four articles were refenced in most of the studies analysed, and these articles were included in the current research, bringing the total number of papers to **22**, shown in **Table 2**, retained for further analysed to develop the measurement tool for CSA.

Table 2: List of publications for CSA model development

N°	Publication	Reference
1	Improving Security Awareness in the Government Sector	(Amjad <i>et al.</i> , 2016)
2	Securing Our Digital Natives: A Study of Commonly Experience Internet Safety Issues and a One-Stop Solution	(Agarwal and Singhal, 2017)
3	Cybersecurity workforce in railway: its maturity and awareness	(Kour and Karim, 2020)
4	Information security awareness and behavior: a theory-based literature review	(Lebek <i>et al.</i> , 2014)
5	Measurement of Employee Information Security Awareness Using the Human Aspects of Information Security Questionnaire (HAIS-Q): Case Study at PT. PQS	(Zulfia <i>et al.</i> , 2019)
6	Measuring the Information Security Awareness Level of Government Employees Through Phishing Assessment	(Ikhsan and Ramli, 2019)
7	Measurement of Employee Information Security Awareness: Case Study at A Government Institution	(Puspitaningrum <i>et al.</i> , 2018)
8	A quick cybersecurity wellness evaluation framework for critical organizations	(Jazri and Jat, 2017)
9	Measuring Information Security Awareness on Employee Using HAIS-Q: Case Study at XYZ Firm	(Cindana and Ruldeviyani, 2019)
10	Measurement of Employee Information Security Awareness Using Analytic Hierarchy Process (AHP): A Case Study of Foreign Affairs Ministry	(Normandia <i>et al.</i> , 2019)
11	A cybersecurity culture research philosophy and approach to develop a valid and reliable measuring instrument	(Da Veiga, 2016)
12	Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q)	(Parsons, McCormac, Butavicius, <i>et al.</i> , 2014)
13	The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies	(Parsons <i>et al.</i> , 2017)
14	Evaluating cybersecurity attitudes and behaviors in Portuguese healthcare institutions	(Nunes, Antunes and Silva, 2021)
15	Information Security Behavior: Development of a Measurement Instrument Based on the Self-determination Theory	(Gangire, Da Veiga and Herselman, 2020)

N°	Publication	Reference
16	Semi-automated Information Security Risk Assessment Framework for Analyzing Enterprises Security Maturity Level	(Abazi and Kő, 2019)
17	Cyber Security Awareness, Knowledge and Behavior: A Comparative Study	(Zwilling <i>et al.</i> , 2020)
18	A Model of Information Security Awareness for Assessing Information Security Risk for Emerging Technologies	(Mejias and Balthazard, 2014)
19	A prototype for assessing information security awareness	(Kruger and Kearney, 2006)
20	A study of information security awareness in Australian government organisations	(Parsons, McCormac, Pattinson, <i>et al.</i> , 2014)
21	Analysis of personal information security behaviour	(Ölütçü <i>et al.</i> , 2016)
22	Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours.	(Hadlington, 2017)

5. Development of CSA Measurement Instrument

This section details the analysis of the articles selected to construct an inclusive CSA measurement tool for the assessment of CSA levels in organisations of interest. Existing models were analysed for similarities, differences, advantages and disadvantages, and a coherent model is proposed based on the analysis.

5.1 Phase 3: Comparison of measurement theories

In phase 3, the selected articles were analysed to identify which base theory was utilised to measure the CSA of the employees in the study. **Table 3** summarises the top base theories used in the selected studies.

Table 3: Summary of recurrence of base theory

N°	Base theory	Utilised in selected articles
1	Knowledge, Attitude, Behaviour model (KAB)	10
2	Human Aspect of Information Security-Questionnaire (HAIS-Q)	8
3	Risky cybersecurity behaviour scale (RScB)	2
4	Attitude Towards Cybersecurity and Cybercrime in Business (ATC-IB)	2
5	ISO 27001	2
6	General deterrence theory (GDT)	2

One of the most popular models on which ten of CSA measurement models in the retained literature were developed was borrowed from the field of social psychology (Kruger and Kearney, 2006). This model is called the Knowledge, Attitude, Behaviour (KAB) Model. Parsons, McCormac, Butavicius *et al.* (2014), Parsons, McCormac, Pattinson *et al.* (2014) and Parsons *et al.* (2017) developed the Human Aspect of Information Security-Questionnaire (HAIS-Q) which was based on the KAB Model discussed by Kearney and Kruger (2006). The HAIS-Q model itself appeared in eight different studies. Other models that appeared in the publications included: Risky cybersecurity behaviour scale (RScB), Attitude Towards Cybersecurity and Cybercrime in Business (ATC-IB), ISO 27001 and General deterrence theory (GDT). Each model had varying areas on which they focused with regards to measuring the CSA of employees. The focus areas for each model were counted and the top recurring focus areas throughout the analysed studies determined. **Table 4** shows a summary of how many times each entry was measured in the selected publications.

Table 4: Summary of number of articles that measured this focus area

N°	Focus area	Times measured in selected articles
1	Internet use	15
2	Password management	14
3	Mobile devices	14
4	Email use	13
5	Social media use	13
6	Incident reporting	12
7	Information handling	10
8	IS policy adherence	6

The selected studies also focused on measurement certain traits possessed by the subjects for which CSA had to be assessed. The authors focused on different traits in order to determine what the level of CSA was for a given set of employees in each organisation evaluated. The most recurring traits were the knowledge, attitude, and

behaviour of the employees. The selected studies focused mainly on 3 types of institutions: public organisations, private organisations and academic institutions.

5.2 Phase 4: CSA Model design

Based on the comparison of the different models, the most popular method relied upon by the majority of the authors the HAIS-Q model, which is based on the KAB model, developed by Kruger and Kearney (2006). They developed a measurement model pegged to three dimensions pertaining to human cognition: knowledge (what you know), attitude (what you think), and behaviour (what you do) (KAB), which each had varying importance levels according to the study. The model developed by Kruger and Kearney (2006) followed strict guidelines of sustainability, ease of use, use of scientific methods, and compliance to the organisation of interest's unique requirement. The approach used by Parsons et al. (2017) in each publication was slightly improved with each iteration. Parsons, McCormac, Butavicius, et al. (2014) outlined the development of HAIS-Q and its connection to the KAB model in their 2014 publication after running a test trial with 500 Australian Employees. The model identified seven focus areas : password management, email use, internet use, social networking sites, mobile computing, information handling, and incident reporting—as opposed to the six proposed by Kruger and Kearney (2006) with social media not being an area of interest at the time. The wide use of HAIS-Q which has also been referenced in numerous publications geared towards measuring the CSA levels in a myriad of organisations makes HAIS-Q a reliable model for further study. Puspitaningrum et al. (2018), Normandia et al., (2019), Cindana and Ruldeviyani (2019), (Zulfia *et al.*, 2019) all based their measurement instruments on the HAIS-Q model. They assessed varying areas and types of organisations from government to private sector employees. The KAB and HAIS-Q models were therefore selected based on popularity with other authors.

5.2.1 Focus areas selection

The selection of the focus areas was done by popularity (see **Table 5**). These focus areas were selected because of how many times they were measured by previous authors. **Table 5** shows which focus areas are common to the different studies and how the current study was developed based on the focus areas previously studied by the experts.

Table 5: Similarities of focus areas in KAB, HAIS-Q and current study

KAB Model	HAIS-Q model	Current study
Focus areas		
Adhere to policies	-	Adhere to policies
Action -> Consequences	-	-
Password management	Password management	Password management
Email use	Email use	Email use
Internet use	Internet use	Internet use
Mobile computing	Mobile computing	Mobile computing
Incident reporting	Incident reporting	Incident reporting
-	Information handling	Information handling
-	Social networking sites	Social networking sites

5.2.2 Key measurement traits selection

Similar to the selection of focus areas, the measurement traits were also selected based on popularity with previous academic publications by authors in the field of interest. The most popular measurement traits after the data extraction process were: Knowledge, Attitude, and Behaviour.

5.3 Model development

The knowledge and attitude aspects of the questionnaire remained generally unaltered, following the traditional methodology of the HAIS-Q model. For this section, respondents will rate the items from 1-5 (where 1 = Strongly Disagree and Strongly Agree) However, items under the behavioural segment of the questionnaire followed the methodology applied in the RScB model, where respondents answered the questions based on retrospective behaviour in the 6 months that precede the evaluation. This was done because the author considers retrospective behaviour more indicative of the respondent's level of awareness than simply having to agree or disagree on what the right behaviour should be in the different instances. The behavioural aspect of the HAIS-Q model was modified to fit RScB model and focus more on retrospective behaviour. The respondents will rate on a scale of 1-5 (where 1 = Never and 5 = Daily or almost daily), how often they engaged in a particular behaviour 6 month prior to the evaluation. The model covers all aspects discussed in the present study and can be used to assess the CSA levels of employees in the water sector by collecting data based on the response of each

individual. A total of 22 questions in 8 focus areas (shown in **Table 5**) forms the final measurement model. A summary of the measurement model, with selected questions from each focus area, can be seen in **Appendix A**).

6. Conclusion

This study documents the process of the development of a model to test cybersecurity awareness in the water sector. This study was based on generic measurement instruments which were not specific to the water sector (although an effort was made to draw parallels between both test cases). Future work includes the utilisation of this measurement model to determine the CSA levels of employees in this sector.

Considering how important CSA is in the water industry, a reliable CSA measurement tool can greatly assist in determining the level of CSA of employees in the sector and assist in the improvement of the security of the systems in this sector. From this study, it can be recommended organisations should invest more resources into evaluating the level of CSA of their employees, especially in the water sector because employees can prove to be the point of entry for some of the most disastrous attack as demonstrated by Hassanzadeh *et al.* (2020). This study also recommends that after performing a diagnostic on the level of CSA of employees, training programs focusing on all areas of weakness should be organised to improve the general level of CSA amongst employees and management.

References

- Abazi, B. and Kő, A. (2019) Semi-automated Information Security Risk Assessment Framework for Analyzing Enterprises Security Maturity Level, Lecture Notes in Business Information Processing. Springer International Publishing. doi: 10.1007/978-3-030-37632-1_13.
- Agarwal, C. and Singhal, A. (2017) 'Securing our digital natives: A study of commonly experience internet safety issues and a one-stop solution', ACM International Conference Proceeding Series, Part F1280, pp. 178–186. doi: 10.1145/3047273.3047303.
- Amjad, H. A. R., Zaffar, M. F., Naeem, U., Choo, K. K. R. and Zaffar, M. A. (2016) 'Improving security awareness in the government sector', ACM International Conference Proceeding Series, 08-10-June, pp. 1–7. doi: 10.1145/2912160.2912186.
- Cindana, A. and Ruldeviyani, Y. (2019) 'Measuring information security awareness on employee using HAIS-Q: Case study at XYZ firm', in 2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018. Institute of Electrical and Electronics Engineers Inc., pp. 289–294. doi: 10.1109/ICACSIS.2018.8618219.
- Gangire, Y., Da Veiga, A. and Herselman, M. (2020) Information Security Behavior: Development of a Measurement Instrument Based on the Self-determination Theory, IFIP Advances in Information and Communication Technology. Springer International Publishing. doi: 10.1007/978-3-030-57404-8_12.
- Hadlington, L. (2017) 'Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours.', Heliyon, 3. Available at: <https://reader.elsevier.com/reader/sd/pii/S2405844017309982?token=825F4D5FD0127B0A8C8C470FC3E81D1ABB4BC845138F39113801012ACC189AF4CB663406D5E4486B841DB671AD4F3925>.
- Hasan, S., Ali, M., Kurnia, S. and Thurasamy, R. (2021) 'Evaluating the cyber security readiness of organizations and its influence on performance', Journal of Information Security and Applications, 58, p. 102726. doi: 10.1016/j.jisa.2020.102726.
- Hassanzadeh, A., Rasekh, A., Galelli, S., Aghashahi, M., Taormina, R., Ostfeld, A. and Banks, M. K. (2020) 'A review of cybersecurity incidents in the water sector', Journal of Environmental Engineering, 146. doi: 10.1061/(asce)ee.1943-7870.0001686.
- Ikhsan, M. G. and Ramli, K. (2019) 'Measuring the Information Security Awareness Level of Government Employees Through Phishing Assessment', 34th International Technical Conference on Circuits/Systems, Computers and Communications, ITC-CSCC 2019, pp. 16–19. doi: 10.1109/ITC-CSCC.2019.8793292.
- Jazri, H. and Jat, D. S. (2017) 'A quick cybersecurity wellness evaluation framework for critical organizations', Proceedings of 2016 International Conference on ICT in Business, Industry, and Government, ICTBIG 2016. doi: 10.1109/ICTBIG.2016.7892725.
- Kong, H. K., Kim, T. S. and Kim, J. (2012) 'An analysis on effects of information security investments: A BSC perspective', Journal of Intelligent Manufacturing, pp. 941–953. doi: 10.1007/s10845-010-0402-7.
- Kour, R. and Karim, R. (2020) 'Cybersecurity workforce in railway: its maturity and awareness', Journal of Quality in Maintenance Engineering. doi: 10.1108/JQME-07-2020-0059.
- Kruger, H. A. and Kearney, W. D. (2006) 'A prototype for assessing information security awareness', Computers and Security, 25(4), pp. 289–296. doi: 10.1016/j.cose.2006.02.008.
- Lebek, B., Uffen, J., Neumann, M., Hohler, B. and Breitner, H. M. (2014) 'Information security awareness and behavior: a theory-based literature review', Management Research Review, 37(12), pp. 1049–1092.

- Luijff, E. (2008) TNO report SCADA Security Good Practices for the Drinking Water Sector TNO | Knowledge for business SCADA Security Good Practices for.
- Luijff, E., Ali, M. and Zielstra, A. (2011) 'Assessing and improving SCADA security in the Dutch drinking water sector', *International Journal of Critical Infrastructure Protection*, 4(3–4), pp. 124–134. doi: 10.1016/j.ijcip.2011.08.002.
- Meades, P. (2015) Doing a literature review in health and social care: a practical guide, *British Journal of Guidance & Counselling*. doi: 10.1080/03069885.2014.975101.
- Mejias, R. J. and Balthazard, P. A. (2014) 'A Model of Information Security Awareness for Assessing Information Security Risk for Emerging Technologies', *Journal of Information Privacy and Security*, 10(4), pp. 160–185. doi: 10.1080/15536548.2014.974407.
- Normandia, Y., Kumaralalita, L., Hidayanto, A. N., Nugroho, W. S. and Shihab, M. R. (2019) 'Measurement of employee information security awareness using analytic hierarchy process (AHP): A case study of foreign affairs ministry', *Proceedings - 2018 4th International Conference on Computing, Engineering, and Design, ICCED 2018*, pp. 52–56. doi: 10.1109/ICCED.2018.00020.
- Nunes, P., Antunes, M. and Silva, C. (2021) 'Evaluating cybersecurity attitudes and behaviors in Portuguese healthcare institutions', *Procedia Computer Science*, 181(2019), pp. 173–181. doi: 10.1016/j.procs.2021.01.118.
- Ölütü, G., Testik, Ö. M. and Chouseinoglou, O. (2016) 'Analysis of personal information security behavior and awareness', *Computers and Security*, 56, pp. 83–93. doi: 10.1016/j.cose.2015.10.002.
- Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A. and Zwaans, T. (2017) 'The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies', *Computers and Security*, 66, pp. 40–51. doi: 10.1016/j.cose.2017.01.004.
- Parsons, K., McCormac, A., Butavicius, M., Pattinson, M. and Jerram, C. (2014) 'Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q)', *Computers and Security*, 42, pp. 165–176. doi: 10.1016/j.cose.2013.12.003.
- Parsons, K., McCormac, A., Pattinson, M., Butavicius, M. and Jerram, C. (2014) 'A study of information security awareness in Australian government organisations', *Information Management and Computer Security*, 22(4), pp. 334–345. doi: 10.1108/IMCS-10-2013-0078.
- Puspitaningrum, E. A., Devani, F. T., Putri, V. Q., Hidayanto, A. N., Solikin and Hapsari, I. C. (2018) 'Measurement of employee information security awareness: Case study at a government institution', *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*. doi: 10.1109/IAC.2018.8780571.
- Rantapelkonen, J. and Salminen, M. (2013) 'The Fog of Cyber Defence', *Department of Leadership and Military Pedagogy Publication*, 2(10), p. 236. Available at: [http://www.doria.fi/bitstream/handle/10024/88689/The Fog of Cyber Defence NDU 2013.pdf?sequence=1&isAllowed=y](http://www.doria.fi/bitstream/handle/10024/88689/The_Fog_of_Cyber_Defence_NDU_2013.pdf?sequence=1&isAllowed=y).
- Snyder, H. (2019) 'Literature review as a research methodology: An overview and guidelines', *Journal of Business Research*, 104(August), pp. 333–339. doi: 10.1016/j.jbusres.2019.07.039.
- Strupczewski, G. (2021) 'Defining cyber risk', *Safety Science*, 135. doi: 10.1016/j.ssci.2020.105143.
- The White House (2013) 'Presidential Policy Directive Critical Infrastructure Security and Resilience.', Technical report.
- Da Veiga, A. (2016) 'A cybersecurity culture research philosophy and approach to develop a valid and reliable measuring instrument', *Proceedings of 2016 SAI Computing Conference, SAI 2016*, pp. 1006–1015. doi: 10.1109/SAI.2016.7556102.
- Zulfia, A., Adawiyah, R., Hidayanto, A. N. and Fitriah Ayuning Budi, N. (2019) 'Measurement of Employee Information Security Awareness Using the Human Aspects of Information Security Questionnaire (HAIS-Q): Case Study at PT. PQS', *5th International Conference on Computing Engineering and Design, ICCED 2019*. doi: 10.1109/ICCED46541.2019.9161120.
- Zwilling, M., Klien, G., Lesjak, D., Wiechetek, Ł., Cetin, F. and Basim, H. N. (2020) 'Cyber Security Awareness, Knowledge and Behavior: A Comparative Study', *Journal of Computer Information Systems*, 00(00), pp. 1–16. doi: 10.1080/08874417.2020.1712269.

Appendix A: Summary of measurement model

Sub Focus Area	Knowledge	Attitude	Behaviour	Behaviour (RScB modification)
Focus Area: IS Policy adherence				
Adhering to IS Policy	It's acceptable to ignore IS protocol set out by management and IT department. *	No serious consequence can result from not following safety protocol at all times at work.	None	Ignores safety protocols because they are difficult or inconveniencing to follow at all times.
Focus Area: Password Management				
Using the same password	It's acceptable to use my social media password on my work accounts. *	It's safe to use same password for social media and work accounts. *	I use a different password for my social media and work accounts.	How often you use your social media password on your work accounts?
Focus Area: Email Use				
Clicking on links in emails from known senders	I am allowed to click on any links in the emails from people I know. *	It's always safe to click on links in emails from people I know. *	I don't always click on links in emails just because they come from someone just because they come from someone I know.	Clicked on a link in an email that came from someone you know without appropriate verification
Focus Area: Internet use				
Downloading files	I am allowed to download any files onto my work computer if they help me to do my work. ^	It can be risky to download files on my work computer.	I download any files onto my work computer that will help me get the job done. *	Downloaded files you considered necessary for work on your work computer.
Focus Area: Social media use				
SM Privacy settings	I must periodically review the privacy settings on my social media accounts.	It's a good idea to regularly review my social media privacy settings.	I don't regularly review my social media privacy settings. *	Reviewed your social media privacy settings.
Focus Area: Mobile device				
Physically securing mobile device	When working in a public place, I have to keep my laptop with me at all times.	When working in a café, it's safe to leave my laptop unattended for a minute. *	When working in a public place, I leave my laptop unattended.	Left your laptop unattended in a public place.
Focus Area: Information handling				
Disposing of sensitive print-outs	Sensitive print-outs can be disposed of in the same way as non-sensitive ones. *	Disposing of sensitive print-outs by putting them in the rubbish bin is safe. *	When sensitive print-outs need to be disposed of, I ensure that they are shredded or destroyed.	Shredded or destroyed sensitive print-outs before disposing of them.
Focus Area: Incident reporting				
Reporting suspicious behaviour	If I see someone acting suspiciously in my workplace, I should report it.	If I ignore someone acting suspiciously in my workplace, nothing bad can happen. *	If I saw someone acting suspiciously in my workplace, I would do something about it.	Intervened after noticing someone acting suspiciously at your work place.

(*Reverse scoring should be used for this item)

Circuit-Variant Moving Target Defense for Side-Channel Attacks

Tristen Mullins, Brandon Baggett, Todd R. Andel and J. Todd McDonald

University of South Alabama, Mobile, USA

tah1323@jagmail.southalabama.edu

brb1323@jagmail.southalabama.edu

tandel@southalabama.edu

jtmcdonald@southalabama.edu

Abstract: The security of cryptosystems involves preventing an attacker's ability to obtain information about plaintext. Traditionally, this has been done by prioritizing secrecy of the key through complex key selection and secure key exchange. With the emergence of side-channel analysis (SCA) attacks, bits of a secret key may be derived by correlating key values with physical properties of cryptographic process execution. Information such as power consumption and electromagnetic (EM) radiation side-channel properties can be observed during encryption or decryption. These signals reflect data-dependent system behaviours that may reveal secret key information. Power and EM SCA attacks require several measurements of the target process to amplify the signal of interest, filter out noise, and derive the secret key through statistical analysis methods. Differential power and EM analysis attacks rely on correlating actual side-channel measurements to hypothetical models. The goal of this research is to increase the complexity of both power and EM SCA by introducing structural and spatial randomization of the target hardware. We propose a System-on-a-Chip (SOC) countermeasure that will periodically reconfigure an AES scheme using randomly located S-box circuit variants. We hypothesize that changing the location of the target modules between encryption runs will result in a nonconstant EM signal strength for any given point on the chip, increasing the number of traces needed to perform a localized EM SCA attack. Further, each of the S-box circuit variants will consist of functionally equivalent, structurally diverse hardware. By diversifying the implementations at the gate-level, we aim to vary the power behaviour observed by the attacker and disrupt the correlation between the hypothetical and actual power consumption, increasing the complexity of power SCA. This moving target defense aims to disrupt side-channel collection and correlation needed to successfully implement an attack.

Keywords: Side-Channel Analysis, Electromagnetic Analysis, Power Analysis, Countermeasure, Reconfigurable Hardware

1. Introduction

Several factors are considered when deciding on which platform to implement a cryptographic algorithm. There are many trade-offs between software and hardware implementations including cost, speed, and flexibility. While software implementations are often more flexible, easier to update, and have low development costs, they can have greater overhead costs and weaker security than their hardware counterparts. Reconfigurable hardware devices (e.g., field-programmable gate arrays, or FPGAs) feature characteristics that allow them comparable flexibility to software implementations while incorporating the benefits of hardware realization. Wollinger and Paar (2003) list some potential advantages reconfigurable hardware provides for cryptography including algorithm agility, algorithm upload, architecture efficiency, resource efficiency, algorithm modification, throughput, and cost efficiency. Not only do these improve algorithm performance, but they also ensure that the platform resources are used efficiently, and updates are easily made through reconfiguration. However, these advantages can only be exploited if security shortcomings are addressed.

The security of cryptosystems involves preventing an attacker's ability to obtain information about plaintext. Traditionally, this has been done by prioritizing secrecy of the key through complex key selection and secure key exchange (Meadows, 2003). With the emergence of side-channel analysis (SCA) attacks, bits of a secret key may be derived by correlating key values with physical properties of cryptographic process execution. Information such as timing (Kocher, 1995), power (Kocher *et al.*, 1999a), and electromagnetic (EM) radiation (Quisquater and Samyde, 2001) side-channel properties can all be observed during run-time of a cryptoprocess. These signals reflect data-dependent system behaviours that may be analysed by an attacker to derive secret key values.

The ability to obtain information about the system is dependent on the accessibility of a “usable” side-channel and does not “reflect inherent weaknesses” of the process being examined (Kelsey *et al.*, 1998). Therefore, countermeasures for SCA attacks should focus on reducing trace usability by minimizing behaviour-key correlation and information leakage within the signal.

Side-channel countermeasures are designed to increase the complexity of SCA attacks. This is often done through hiding and masking techniques such as random delay insertion, shuffling, masking, dual-rail logic, etc.

Increasing attack complexity does not necessarily mean that an attack is impossible, just more costly. Because there is currently no solution that eliminates all side-channel leakage, we must “accept that cryptographic implementations leak a certain amount of information,” and avoid allowing the leakages to completely compromise security during use (Standaert *et al.*, 2010).

Many researchers suggest implementing countermeasures in combination to compensate for the shortcomings of individual designs and improve overall security (Güneysu and Moradi, 2011; Moradi, Mischke and Paar, 2011). This method of layering side-channel resistance may also be useful in scenarios where an attacker could perform multiple types of SCA on a target. With physical access to a device, both power and electromagnetic analysis (EMA) attacks may be conducted with simple equipment. Though many power countermeasures are assumed to protect against EMA, it has been shown that power countermeasures may still be vulnerable to localized EMA attacks (Agrawal *et al.*, 2002; Immler, Specht and Unterstein, 2017; Specht *et al.*, 2018). This creates a need for both power and EMA attack prevention methods on a device.

This research focuses on increasing the complexity of both power and localized EM SCA by introducing structural and spatial randomization of the target hardware. The proposed design utilizes the device’s dynamic reconfiguration capabilities to enable a varying attack surface for a parallelized version of AES-128. Our countermeasure intends to reduce the usable trace set for an attack by incorporating circuit variants of the encryption hardware and limit EM hotspots by randomizing the hardware location.

The remainder of this paper is organized as follows. In Section 2, we outline the side-channel attacks our countermeasure is designed to prevent including differential power analysis, localized EM analysis, and exploitable leakage present in AES. Our proposed countermeasure is described in Section 3 and related works are presented in Section 4. The paper concludes in Section 5.

2. Side-Channel Analysis

2.1 Power Side-Channel Analysis

Kocher *et al.* (1998) first demonstrated how power consumption measurements could be correlated to secret key values. These power analysis attacks are based the behaviour of semiconductor logic. When charge is applied to or removed from transistors, a current is induced that consumes power and emits EM radiation. This switching activity may vary depending on which operations are being executed and even the data values being processed. These trends in power consumption may be measured by an attacker and used to determine runtime information that may otherwise be assumed to be private. If the device under observation is executing cryptographic processes, the data-dependent power usage may expose the secret key.

Differential power analysis (DPA) attacks use algorithm-specific statistical methods to identify data-dependent correlations in power traces. This type of attack requires little knowledge about the target device, requiring only the target cryptographic algorithm to be known. The strength of DPA attacks depends on the power model the attacker chooses. Hamming-distance and Hamming-weight models are the most common power models used in DPA attacks due to their ease of application (Mangard *et al.*, 2007). Customized power models increase the effectiveness of the attack but are up to the attacker to derive using their knowledge of the device.

DPA attacks all follow a general 5-step procedure (Mangard *et al.*, 2007). In step 1, an intermediate value is selected on which the attack will be based. This value must be a function of the plaintext and the key. Step 2 consists of measuring actual power traces from the target device. In step 3, a set of hypothetical intermediate values are calculated using a set of key hypotheses for the target cryptographic algorithm. The values from step 3 are then mapped to their hypothetical power consumption using the power model the attacker has selected as step 4. The hypothetical power consumption is then compared to the actual power consumption in step 5. The attacker assumes there is 1 hypothetical power consumption “trace” that corresponds to the correct key and 1 time period in the actual trace set where the selected intermediate value was processed. High correlation between hypothetical and actual power consumption values are indicative of a correct key hypothesis.

For DPA to be mounted efficiently, the captured power traces must be aligned in time (Mangard *et al.*, 2007). This ensures that each cryptographic operation corresponds to the same location throughout the trace set which is necessary for comparing the hypothetical power consumption to the actual trace set. Several countermeasures have been shown to be effective against DPA by misaligning traces including operation

shuffling, insertion of dummy rounds, and random delay insertion (Clavier, Coron and Dabbous, 2000; Korak, Plos and Hutter, 2012; Bogdanov *et al.*, 2019). These types of countermeasures increase the cost of performing an attack by reducing the number of usable traces that can be collected in a given amount of time.

2.2 Electromagnetic Side-Channel Analysis

While data-dependent current flow serves as the basis for power analysis attacks, it also emits electromagnetic fields that can contain key-revealing information. Differential Electromagnetic Analysis (DEMA) attacks follow similar statistical analysis methods to DPA using signals collected from EM field probes (Kocher *et al.*, 2011, 1999b; Quisquater & Samyde, 2001). Using EM rather than current-based power measurements for attacks does have its advantages. EM measurements offer a desirable alternative to power consumption when access to the power and ground lines are limited, when the power signal contains too much noise, or when power analysis countermeasures are implemented (DeBeer *et al.*, 2011; Gandolfi *et al.*, 2001; Quisquater & Samyde, 2001).

When using a high-resolution EM probe, attackers may perform localized attacks using traces collected from a specific area on the chip (Gandolfi, Mourtel and Olivier, 2001). Localized EM attacks are particularly powerful in that they allow leakage from individual subcircuits to be isolated (Li, Iyer and Orshansky, 2019a). This may be leveraged by attackers to weaken countermeasures that are effective against DPA. Specht *et al.* (2018) use localized EMA to isolate the leakage from separate shares in a threshold implementation countermeasure. Their attack combines leakage from multiple probes to break the scheme. Localized EM attacks have also been used to break dual-rail logic countermeasures by leveraging placement and routing imbalances (Immler, Specht and Unterstein, 2017).

These types of attacks are most successful when the probe is placed above the area of the chip where the side-channel leakage of interest is strongest (Heyszl *et al.*, 2012). To determine the optimal probe position, measurements are taken across the surface of the chip during the execution of the target process. If the chip hosts a variety of processes with distinct clock frequencies, the EM signal should be filtered to reduce components that are not related to the target operations. The location with the greatest signal strength for the target process or the highest correlation to the hypothetical model indicates the optimal probe position. These points of interest may be visualized as “hotspots” using a spectral intensity plot or a correlation heat map. Once a target hotspot is selected, the probe is then placed, and an attack trace set is collected. Like the traces used for DPA, traces used for DEMA also need to be aligned during pre-processing.

2.3 Side-Channel Leakage in AES

The Advanced Encryption Standard (AES) is the current standard for encrypting electronic data (Standard, 2001). This symmetric block cipher is a form of the Rijndael cipher (Daemen and Rijmen, 1998) that processes 128-bit blocks with variable key length.

Figure 1 shows a block diagram for AES encryption. After the initial round key addition, a round function is implemented either 10, 12, or 14 times depending on the key length. Each round, excluding the final, consists of four transformations: SubBytes, ShiftRows, MixColumns, and AddRoundKey. For a parallelized version of AES, 16 copies of the S-box are used so that all 16 bytes of a data block may be substituted at once.

The SubBytes transformation is a non-linear byte substitution that operates on each byte independently using a substitution table, or S-box. S-boxes are of particular interest to attackers due to their output being a function of the plaintext and the key in the first round of AES (Mangard *et al.*, 2007). Therefore, this value is often the intermediate value on which DPA and DEMA attacks on AES are based. While simple countermeasures may be used to mask the leakage of other intermediate values for the algorithm, the non-linearity of the SubBytes transformation makes it difficult to mask (Oswald *et al.*, 2005). Masked S-box implementations may still leak information via glitches when realized in hardware, requiring the inclusion of additional countermeasures (Mangard and Schramm, 2006). Localized EM attacks are also a significant threat to AES implementations since they allow leakage from a particular S-box to be isolated (Unterstein *et al.*, 2017).

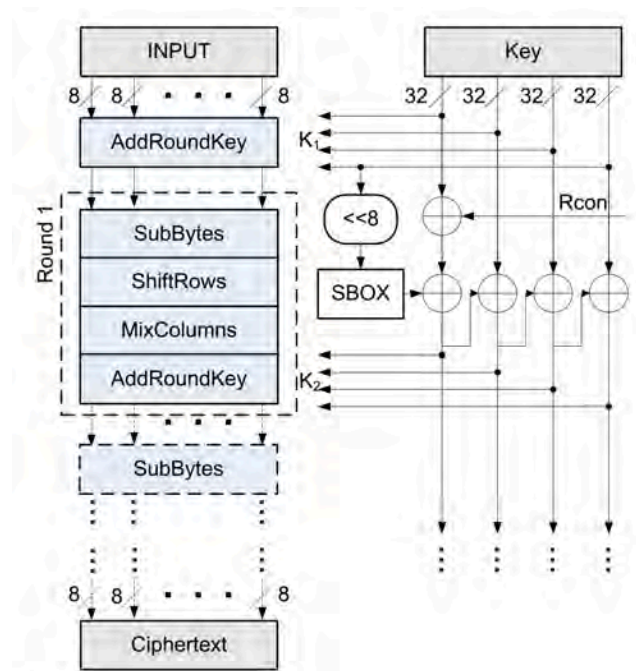


Figure 1: AES Encryption Block Diagram (Ambrose, Parameswaran and Ignjatovic, 2008)

3. Proposed Design

3.1 Equipment and Resources

This research will use a Digilent ZedBoard evaluation and development platform, which features a Zynq-7000 SoC (Xilinx, 2019). To design and program the system, we will use Xilinx Vivado Design Suite. With this software, we can synthesize designs from behavioural descriptions (e.g., VHDL code), add and configure specialized IP cores, specify placement and route (P&R) details, simulate execution, and generate then export bitstreams to the device. A Java software package called the Program Encryption Toolkit (PET) (Forbes, 2017) will be used to generate circuit variants of AES S-box structures that will be used to modify the design in Vivado.

To collect power and EM traces for our analysis we will use a Riscure Side-Channel analysis Suite including an EM Probe station, PicoScope 3000 Series oscilloscope, and Inspector software. The probe station consists of a high-resolution EM probe and a motorized XYZ table which are integrated with the Inspector software for configuration and measurement. The station can be setup to automatically scan the surface of the chip with a step size as small as 2.5 μm . The Inspector software will be used to configure the equipment for trace collection, send plaintext to the target device, receive ciphertext output, store traces, perform pre-processing on traces (e.g., filtering and resampling), and statistically analyse the samples during the attack.

For this research, we will implement a parallelized hardware version of AES-128. This will limit the algorithm to ten encryption rounds for a 128-bit block of plaintext which will be randomly generated by the Inspector software during analysis.

3.2 System Overview

The goal of this research is to increase the complexity of both power and localized EM SCA by introducing structural and spatial randomization of the target hardware. We propose a countermeasure that will periodically reconfigure an AES scheme using randomly located circuit variants of the S-boxes. The focus of this moving target defense (MTD) approach is limiting the presence of EM “hotspots” that indicate favourable candidates for high-resolution probe placement. Changing the location of the target components between encryption runs will result in a nonconstant signal strength for any given point on the chip, increasing the number of traces needed to perform a localized EMA attack. To implement this feature, an additional IP core will be configured that will serve as a Partial Reconfiguration Controller (PRC). Upon the occurrence of a trigger event, the PRC will fetch a partial bitstream from memory that will be used to reconfigure the FPGA. An interval on which reconfiguration should occur will need to be determined as well as what portions of the programmable logic (PL) will be reserved for reconfigurable area.

Power analysis resistance will be introduced to the design through the partial bitstreams used for reconfiguration. Each of the S-box bitstreams stored in memory will consist of functionally equivalent, structurally varied implementations. By diversifying the implementations at the gate-level, we aim to vary the power behaviour observed by the attacker and disrupt the correlation between the hypothetical and actual power consumption.

A Program Encryption Toolkit (PET) will be used to generate circuit variants for each S-box. This is a customized Java application that was used in the research presented by Forbes et al. (2017). It includes a feature to generate random equivalent circuits based on a ISCAS format netlist. After selecting an original circuit, the user specifies the number of gates, gate types, and maximum fan in for the variant. For the merged signature method, a circuit is generated for each output of the original and then merged to form a single circuit. The resulting circuits may be exported in several formats including BENCH netlist and VHDL. A tool is also included to compare BENCH netlists to verify that they are semantically equivalent. Figure 2 provides an example of the variants generated by PET using merged signature circuit generation.

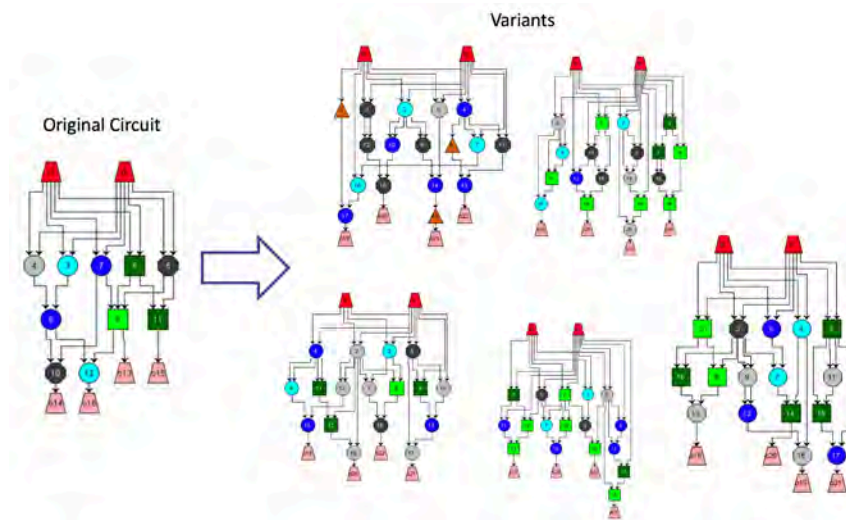


Figure 2: Merged Signature Circuit Generation. 5 Gates with a Max Fan In of 2

A practical countermeasure evaluation will be conducted to determine the implementation's resistance to power and localize EM analysis. DPA and DEMA attacks will be performed on a static, unprotected target as a control. The number of traces to break this AES version will indicate the worst-case resistance to each attack. The DPA and DEMA resistance of the countermeasure will be assessed by comparing the number of traces to key disclosure (NTD) for each attack to that of the unprotected version. During the acquisition step of the DEMA attacks, the optimal probe placement will be determined as well as any leakage hotspots identified. These characteristics will be used to investigate the level of hotspot obfuscation provided by the countermeasure.

Figure 3 shows a block diagram for the proposed design. The Zynq System on a Chip (SOC) features a processing system (PS) and FPGA programmable logic (PL). Though the AES implementation is hardware based, the input and output are handled by the PS side of the SOC. Interfacing applications will be hosted on the PS side that will be used to communicate to the device with the Inspector software (i.e., send and receive data blocks). During each execution of the algorithm, a 128-bit block of random data will be sent from the Inspector software to be encrypted by the AES scheme.

The entire AES core will be designed using a behavioural VHDL description and the Xilinx Vivado software. Excluding the S-boxes, the AES core will be placed in a designated area in the PL. This area is near the reconfiguration region to avoid having to make connections over long distances which can be more prone to leakage. The static AES modules will include a signal that will be used to indicate to the oscilloscope when to start and stop trace collection during analysis. This trigger will correspond to the receiving of plaintext and sending of ciphertext to the PS side of the SOC.

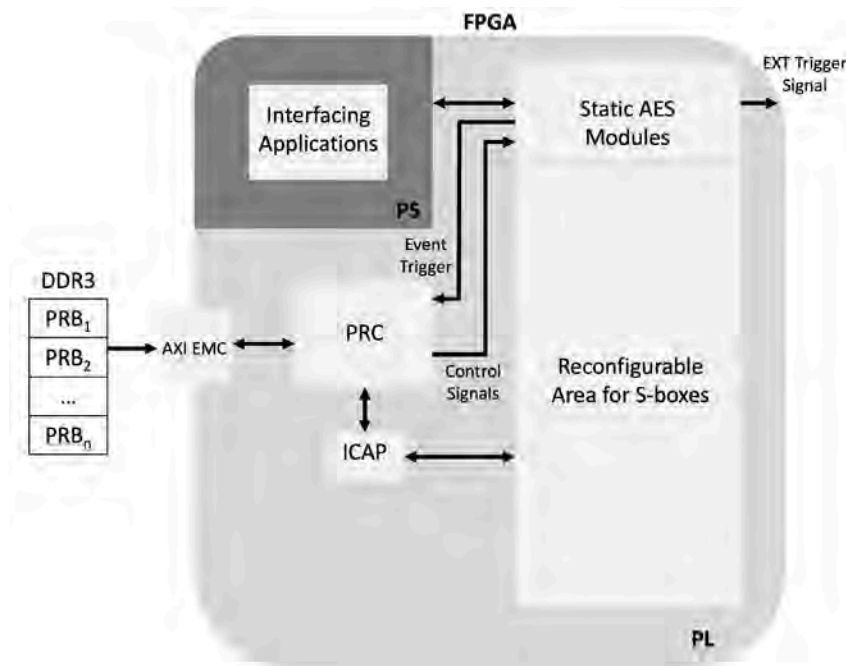


Figure 3: System Block Diagram

The target device supports the inclusion of a Partial Reconfiguration Controller (PRC) IP core (Xilinx, 2018). Upon the occurrence of a trigger event, the PRC fetches a partial bitstream from memory for reconfiguration. In this design, the PRC will be triggered after a certain number of encryptions have been performed. This number will be based on the minimum number of traces needed to break a single configuration. The PRC delivers the PRB to an internal configuration access port (ICAP) which performs the partial reconfiguration. Any control signals needed to disable existing reconfigurable modules or protect static modules are handled by the PRC. Due to the limited size of available BRAM on the FPGA, it is likely that PRBs will need to be stored in external flash memory, DDR3. The PRC can access the PRBs using and AXI External Memory Controller (AXI_EMC).

4. Related Works

Diversity for obfuscation was first presented by Cohen to protect operating systems (1993). In their work, several methods for increasing attack complexity via program evolution are discussed. This concept of implementation diversity has since grown to include schemes for hardware security.

In their work, Li et al. (2019) propose a spatial randomization technique to minimize leakage that may be exploited through localized EM attacks. In this MTD countermeasure, a permutation network is used to randomly assign data bytes to 16 AES S-boxes and a second permutation network restores the order of the bytes. Unlike the design we have proposed which uses dynamic partial reconfiguration (DPR) to randomize the target location, the countermeasure presented by Li et al. utilizes dynamic logic reconfiguration (DLR) which selects among components that remain static in the logic of the FPGA.

Hettwer et al. (2019) demonstrate that randomized hardware location may be used to minimize observable EM leakage hotspots. In their work, the entire AES core is replaced between encryption runs and placed in a different location in the FPGA fabric. The authors also introduce implementation diversity by setting minimum and maximum path delay restrictions to randomize the dynamic power consumption. Their results showed an increase in localized EMA and fault injection resistance; however, resistance to power analysis was only increased by a factor of 3. Additional countermeasures would be needed to further increase resistance to power analysis attacks.

In their work, Bow et al. (2020) present Side-Channel Power Resistance for Encryption Algorithms Using Implementation Diversity (SPREAD). This paper is a power countermeasure proof-of-concept that shows that input-to-output delays are distinct for each circuit structure used. The authors use two methods for varying S-box behavior: synthesis and circuit-directed variation. The synthesis-directed method creates diversity by excluding specific gate types from netlists and the circuit-directed introduces random delays on paths.

This research differs from related works in both the circuit variant generation specifications and the increased resistance to localized EM attacks. By limiting the design to only reconfigure S-boxes, we expect to reduce the storage overhead of the scheme proposed by Hettwer et al. (2019). The criteria for generating variants in our proposal focuses on gate-level diversity as well as P&R constraints which is expected to further increase resistance to power attacks. The bitstreams that will be used in this research include the P&R configurations for the netlist where the SPREAD (2020) scheme utilizes a custom synthesis tool flow to create relocatable bitstreams. Though this may result in less storage overhead for SPREAD, it is likely that our reconfiguration controller will be simpler, leaving more FPGA area to dedicate to the S-box reconfigurations. This research will also only utilize a synthesis-directed circuit variant generation method as opposed to SPREAD which also includes additional hardware for a circuit-directed approach. To introduce gate-level diversity, we will use a program encryption toolkit (Forbes, 2017) to generate equivalent circuits that vary in size and composition where the method in (Bow et al., 2020) exclude specific gate types when generating netlists for each version. Lastly, the proposed countermeasure will be not only be assessed for power analysis resistance but for localized EM analysis resistance as well.

5. Conclusion

Related work has shown that individually, spatial randomization and circuit variance may be used to obfuscate optimal EM probe positions and increase resistance to power analysis attacks, respectively. However, countermeasures that have attempted to combine these concepts have yet to display resistance to both power analysis and localized EMA attacks in an efficient manner. A spatially randomized implementation may hinder an attacker using a high-resolution EM probe, but if the power consumption behaviour does not vary between implementations, an attacker may still perform a side-channel attack that is not location-dependent (i.e., a power analysis attack). This scenario is unfavourable since the equipment to perform a power analysis attack is simpler and more affordable than that of a localized EM attack (Mangard et al., 2007; Heyszl et al., 2012). Therefore, it is in the researcher's best interest to ensure that defenses against localized EM attacks are also resistant to power attacks.

This research proposes a countermeasure design to resist DPA and DEMA attacks by structurally and spatially randomizing a design. It is hypothesized that reconfiguring an AES implementation with S-box circuit variants will result in a randomized power profile that will disrupt the correlation between actual and hypothetical power consumption, hindering DPA attacks. It is also hypothesized that shuffling the location of S-boxes will obfuscate candidates for optimal probe placement, increasing resistance to localized DEMA attacks.

Reconfiguring the S-box area between encryption runs is expected to increase the execution time of our implementation. Further, the inclusion of control logic for reconfiguration is expected to increase the area of the design. This cost as well as the need for external storage for bitstreams may limit the applicability of our countermeasure to other devices that do not have such resources available. Future work will focus on improving these timing costs as well as determining the minimum number of bitstreams needed for power analysis resistance to reduce storage overhead costs.

References

- Aerts, W. et al. (2006) 'Matching shielded loops for cryptographic analysis', in *1st European Conference on Antennas and Propagation-EuCAP 2006*, pp. 1–6.
- Agrawal, D. et al. (2002) 'The EM side—channel (s)', in *International workshop on cryptographic hardware and embedded systems*, pp. 29–45.
- Ambrose, J. A., Parameswaran, S. and Ignjatovic, A. (2008) 'MUTE-AES: A multiprocessor architecture to prevent power analysis based side channel attack of the AES algorithm', in *2008 IEEE/ACM International Conference on Computer-Aided Design*, pp. 678–684.
- Bogdanov, A. et al. (2019) 'Higher-order DCA against standard side-channel countermeasures', in *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pp. 118–141.
- Bow, I. et al. (2020) 'Side-Channel Power Resistance for Encryption Algorithms Using Implementation Diversity', *Cryptography*, 4(2), p. 13.
- Clavier, C., Coron, J.-S. and Dabbous, N. (2000) 'Differential power analysis in the presence of hardware countermeasures', in *International Workshop on Cryptographic Hardware and Embedded Systems*, pp. 252–263.
- Cohen, F. B. (1993) 'Operating system protection through program evolution', *Computers and Security*, 12(6), pp. 565–584. doi: 10.1016/0167-4048(93)90054-9.
- Daemen, J. and Rijmen, V. (1998) 'The block cipher Rijndael', in *International Conference on Smart Card Research and Advanced Applications*, pp. 277–284.

- DeBeer, F. et al. (2011) 'Practical electromagnetic analysis', in *Proc. Non-Invasive Attack Testing Workshop NIAT*.
- Forbes, M. A. (2017) *Digital Logic Protection Using Functional Polymorphism and Topology Hiding*. University of South Alabama.
- Gandolfi, K., Mourtel, C. and Olivier, F. (2001) 'Electromagnetic analysis: Concrete results', in *International workshop on cryptographic hardware and embedded systems*, pp. 251–261.
- Güneysu, T. and Moradi, A. (2011) 'Generic Side-Channel Countermeasures for Reconfigurable Devices', *CHES 2011*, pp. 33–48.
- Hettwer, B. et al. (2019) 'Securing cryptographic circuits by exploiting implementation diversity and partial reconfiguration on FPGAs', in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 260–263.
- Heyszl, J. et al. (2012) 'Strengths and limitations of high-resolution electromagnetic field measurements for side-channel analysis', in *International Conference on Smart Card Research and Advanced Applications*, pp. 248–262.
- Immler, V., Specht, R. and Unterstein, F. (2017) 'Your rails cannot hide from localized EM: how dual-rail logic fails on FPGAs', in *International Conference on Cryptographic Hardware and Embedded Systems*, pp. 403–424.
- Kelsey, J. et al. (1998) 'Side channel cryptanalysis of product ciphers', in *European Symposium on Research in Computer Security*, pp. 97–110.
- Kocher, P. et al. (1998) 'Introduction to differential power analysis and related attacks', 4(1), pp. 64–75.
- Kocher, P. et al. (1999a) 'Differential power analysis', in *Annual International Cryptology Conference*, pp. 388–397. doi: 10.1007/978-1-4419-5906-5_196.
- Kocher, P. et al. (1999b) 'Differential power analysis', in *Annual International Cryptology Conference*, pp. 388–397. doi: 10.1007/978-1-4419-5906-5_196.
- Kocher, P. et al. (2011) 'Introduction to differential power analysis', *Journal of Cryptographic Engineering*, 1(1), pp. 5–27. doi: 10.1007/s13389-011-0006-y.
- Kocher, P. C. (1995) 'Cryptanalysis of Diffie-Hellman, RSA, DSS, and other systems using timing attacks', in *Extended abstract*.
- Korak, T., Plos, T. and Hutter, M. (2012) 'Attacking an AES-enabled NFC tag: Implications from design to a real-world scenario', in *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pp. 17–32.
- Li, G., Iyer, V. and Orshansky, M. (2019a) 'Securing AES against Localized em Attacks through Spatial Randomization of Dataflow', *HOST 2019*, pp. 191–197. doi: 10.1109/HST.2019.8741026.
- Li, G., Iyer, V. and Orshansky, M. (2019b) 'Securing AES against Localized em Attacks through Spatial Randomization of Dataflow', *HOST 2019*, pp. 191–197. doi: 10.1109/HST.2019.8741026.
- Mangard, S. et al. (2007) *Power Analysis Attacks: Revealing the Secrets of Smart Cards (Advances in Information Security), Power Analysis Attacks*. Available at: <http://gen.lib.rus.ec/book/index.php?md5=16648344AC0BC386A97D0ADACC9F9F78>.
- Mangard, S. and Schramm, K. (2006) 'Pinpointing the side-channel leakage of masked AES hardware implementations', *CHES 2006*, 4249 LNCS, pp. 76–90. doi: 10.1007/11894063_7.
- Meadows, C. (2003) 'What makes a cryptographic protocol secure? the evolution of requirements specification in formal cryptographic protocol analysis', in *European Symposium on Programming*, pp. 10–21.
- Moradi, A., Mischke, O. and Paar, C. (2011) 'Practical evaluation of DPA countermeasures on reconfigurable hardware', *HOST 2011*, pp. 154–160. doi: 10.1109/HST.2011.5955014.
- Oswald, E. et al. (2005) 'A side-channel analysis resistant description of the AES S-box', in *International workshop on fast software encryption*, pp. 413–423.
- Quisquater, J.-J. and Samyde, D. (2001) 'Electromagnetic analysis (Ema): Measures and countermeasures for smart cards', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2140, pp. 200–210. doi: 10.1007/3-540-45418-7_17.
- Quisquater, J.-J. and Samyde, D. (2001) 'Electromagnetic analysis (ema): Measures and counter-measures for smart cards', in *International Conference on Research in Smart Cards*, pp. 200–210.
- Specht, R. et al. (2018) 'Dividing the threshold: Multi-probe localized em analysis on threshold implementations', in *2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pp. 33–40.
- Standaert, F.-X. et al. (2010) 'Leakage resilient cryptography in practice', in *Towards Hardware-Intrinsic Security*. Springer, pp. 99–134.
- Standard, N.-F. (2001) 'Announcing the advanced encryption standard (AES)', *Federal Information Processing Standards Publication*, 197(1–51), p. 3.
- Unterstein, F. et al. (2017) 'Dissecting Leakage Resilient PRFs with Multivariate Localized EM Attacks A Practical Security Evaluation on FPGA', *COSADE 2017*, 2, pp. 120–137. doi: 10.1007/978-3-319-64647-3.
- Wollinger, T. and Paar, C. (2003) 'How secure are FPGAs in cryptographic applications?', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2778, pp. 91–100. doi: 10.1007/978-3-540-45234-8_10.
- Xilinx (2018) 'Partial Reconfiguration Controller', *Xilinx*. Available at: https://www.xilinx.com/support/documentation/ip_documentation/prc/v1_3/pg193-partial-reconfiguration-controller.pdf.
- Xilinx (2019) 'Zynq-7000 SoC product selection guide', *Xilinx*. Available at: <https://www.xilinx.com/support/documentation/selection-guides/zynq-7000-product-selection-guide.pdf>.

A Modern ICT Network Simulator for Co-Simulations in Smart Grid Applications

Fabian Niehaus¹, Bastian Fraune², Giacomo Gritzan³ and Richard Sethmann⁴

FRI, Fakultät Elektrotechnik und Informatik, Bremen, Germany*

fri@hs-bremen.de¹

bastian.fraune@hs-bremen.de²

giacomo.gritzan@hs-bremen.de³

sethmann@hs-bremen.de⁴

Abstract: The current transformation of power grids towards smart grids and the associated increase in the use of ICT technologies lead to an increased attack surface via the communication network. Holistic, multi-domain co-simulations can be used to evaluate the possibilities for, and impacts of, ICT-based attacks on these grids. This paper presents an ICT network simulator that can be used in co-simulation environments. It outlines requirements for this kind of simulation tool and provides insight into how these requirements can be met. In addition, co-simulation use cases for the simulator arising from different research projects are laid out. Finally, the simulator's functionalities in co-simulation environments are verified with a practical application example. This application shows that our simulator (rettij) is able to exchange data with other simulators through a co-simulation interface. It also demonstrates our simulators capabilities to perform real-world ICT attacks on realistic network topologies.

Keywords: ICT; Cyber Security; Co-Simulation; Network Simulator; Simulation; Smart Grid; Power Grid; Kubernetes; Docker

1. Introduction

Many power grids are currently undergoing a transformation towards smart grids (Doleski, 2017). Smart grids change the current power system from a centralized production approach to a more flexible and distributed integration of renewable energy resources. This results in an increased intertwining of supervisory control and data acquisition (SCADA) systems and information and communication technology (ICT). The communication infrastructure is one of the key drivers towards such a smart grid, as the command and control of many distributed energy resources needs to be managed and coordinated. Consequently, intensive use of ICT is necessary to transform the power grid. This increases the attack surface on the overall grid (M. Iguere, et al., 2006) (Yan, et al., 2013). As smart grids have strong interconnections with ICT systems, analyzing the impact of ICT-based attacks becomes a complex task. It is therefore necessary to be able to analyze and understand the effects of ICT attacks on the power grid in greater depth. This can be done by co-simulating power systems and their associated communication networks (Tuinema, et al., 2020). An isolated approach is not sufficient from the perspective of IT security since attack vectors between the systems are increasing due to the large number of small, distributed components. Those ICT attacks generate a demand for simulation testbeds which behave as similar as possible to real ICT infrastructures.

As a result of the issues outlined above, the overarching research question of "How can we holistically simulate future smart grid infrastructures?" arises. In this publication, we aim to propose a possible solution to the issue of how to integrate an ICT simulation with simulations of other associated domains. The requirements for such a simulation tool are derived in part from our federally funded research projects, which address smart grids and ICT attacks and are described in section Use Cases in Smart Grid Research Projects. We identified the following requirements for this ICT simulator:

- Seamlessly and natively work within co-simulations.
- Easily integrate custom and well-known components such as virtual remote terminal units (vRTUs) or web servers.
- Display realistic network traffic without noise from management systems such as Docker.
- Be able to use the Linux network stack.
- Integrate easily configurable and human readable network topologies.
- Support distributed execution.
- Integrate the well-known format of Dockerfile-based container images.

As stated by Muller et al., the development of scalable frameworks that combine ICT and power simulation needs to be improved (Müller, et al., 2019). They have identified three mayor challenges to overcome with in

future research: Scalability, reuse of existing work through open source access, and enabling comparative benchmarking. In order to contribute to solving those challenges, we present our network simulator's capabilities to build network infrastructures and the simple integration of containers that can be used to integrate realistic ICT components for scalable co-simulations together with power domain related simulators. The rest of the paper is structured as follows. First, an overview of related works is given in section Related Work. Our ICT simulator is presented in section Presenting our Network Simulator, and an overview over its most important features regarding the challenges outlined above is provided. Section Use Cases in Smart Grid Research Projects shows three use cases for our simulator in the context of current smart grid research projects. To verify our simulator's functionality in a co-simulation environment, a practical application example is presented in section Practical Application Case Study. The paper ends with a summary and outlook in section Conclusion.

2. Related Work

In the recent past, many co-simulation approaches have been published, most of them in the domain of power grids. This can be explained by the increase of communication infrastructure within these grids. The continuously increasing attack surface and complexity of cross-domain integration necessitate the ability to carry out the effects and detection of attacks in simulations. This section presents recent publications that are relevant for our work. To the best of our knowledge, none of these co-simulation-capable approaches are using containers to run software or vRTUs within their simulation scenarios.

In a recent publication, Muller et al. list ten co-simulation setups that include ICT networks in the simulations, and compare them in detail (Müller, et al., 2019). It can be observed that ns-2 (the predecessor of ns-3) and OPNET Modeler are among the most frequently mentioned network simulators in these co-simulation approaches. While both concepts and software artefacts for a co-simulation of the power grid exist, they generally lack scalability. Their usability for third parties not involved with the development of the approaches is also in need of improvement, partly due to them often not being open source.

Moulema et al. define various co-simulation scenarios with ICT networks (Moulema, et al., 2015). These scenarios are used to examine the impact of poor communication network performance as well as IT attacks on the energy market price. The co-simulation was implemented using the Fenix Framework for Network Co-Simulation (FNCS) in conjunction with GridLAB-D and MATPOWER. The communication network has been simulated using ns-3. The co-simulation scenarios consider demand and response during False-Data-Injection Attack (FDIA) as well as simulate a low data rate for the communication between smart meters and their corresponding gateways. The authors have shown that FDIAs have an impact on the energy market by increasing prices. However, it remains unclear how the FDIAs have been simulated in detail. It is also unclear how the market application was integrated into the co-simulation. Such an implementation of the co-simulation seems sufficient for their purpose, but it doesn't seem to be sufficient from the IT security perspective for analyzing and evaluating IT security tools like intrusion detection systems in combination with real-world attacks on smart grids.

Another co-simulation architecture is presented by Mirz et al. (Soares, et al., 2008). It also combines the three domains energy market, power system and communication network. The authors propose a model-based topology builder based on the IEC Common Information Model (CIM). These models are used to describe an entire smart grid topology including the corresponding communication and market entities. Their architecture contains one simulator for each domain which are coordinated by the co-simulation framework mosaik. The power system simulation has been done with Modelica, the market simulation focusing Virtual Power Plants (VPPs) has been implemented in Python and the communications network has been simulated using ns-3. While this seems to be a promising approach with regard to the communication network and the IT security aspects, failures in the communications network have been investigated in only one scenario.

3. Presenting our Network Simulator rettij

In this section, we present our network simulator that we have named rettij¹ and its capabilities regarding these previously identified challenges:

- The simulated network and network-connected devices should display realistic behavior.

¹ Open Source rettij Network Simulator on GitLab: <https://gitlab.com/frihsb/rettij>

- The simulator needs native support for co-simulation interfaces.
- The simulator should be able to run large-scale simulations without the need for costly, specialized hardware.
- Simulated, virtualized and physical domain-specific components should be easily integrable.

Our network simulator aims to provide a solution to these challenges. The basic functionality has already been described by Woltjen, et al. in (Woltjen, et al., 2020). It simulates network-connected devices (*Node*), their network interfaces (*Interface*) and their network connections to other Nodes (*Channel*). These networks are defined via a *Topology*, which describes the components and their connections in a human-readable YAML file format. The Nodes are primarily run as containers, which allows for lightweight, scalable and repeatable execution. Our simulator rettij uses Kubernetes² for the distributed orchestration and network attachment of the containers. Users can use their preferred Kubernetes implementation and container runtime, and even scale the cluster into cloud environments. A simple example for such a simulation network is shown in Figure 1.

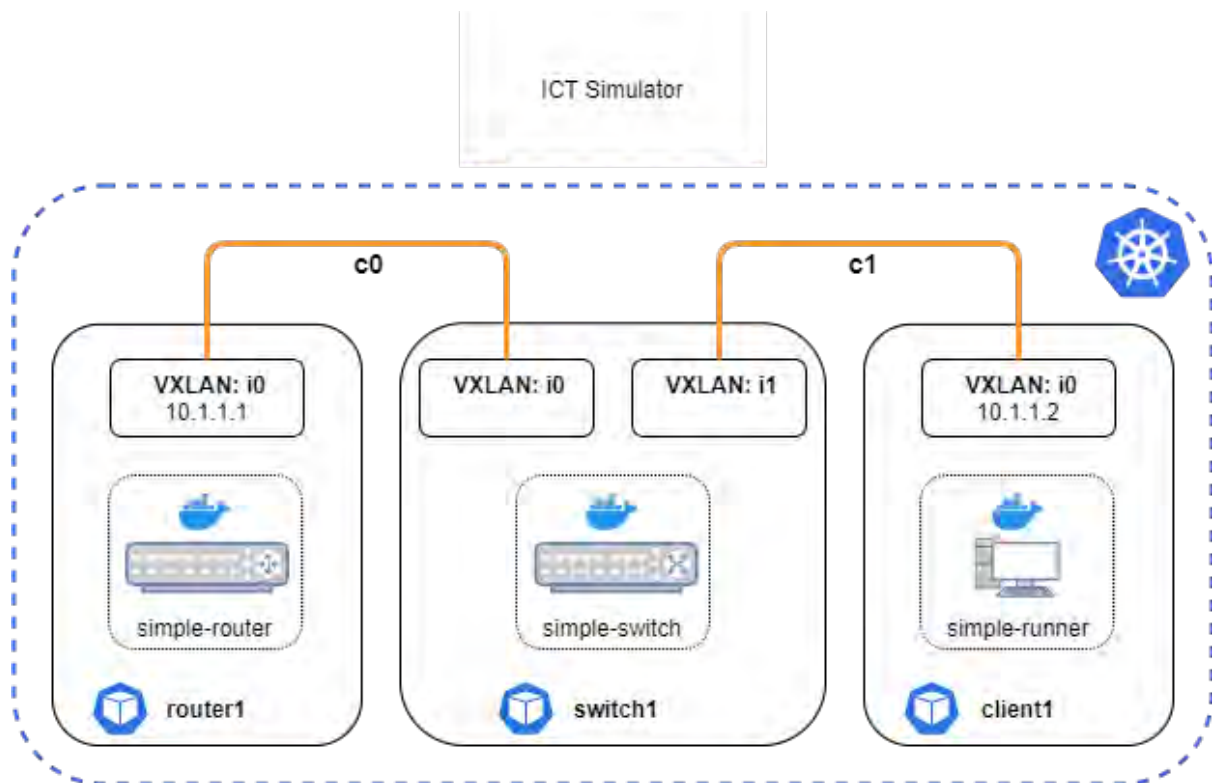


Figure 1: Simple simulation network

As various simulation scenarios require different network devices, our simulator rettij allows the definition of custom components and configurations. These use the Docker Image³ file format in combination with a specification file and hooks to be executed at certain times during the simulation lifecycle. The use of Docker Images enables users to easily integrate commonly used real-world software images from sources like Docker Hub⁴. Users can also build their own images based on a well-documented file format. The simulator ships with pre-defined components for the most important network devices (hub, router, switch) as well as a basic network client.

3.1 Network Encapsulation

One of the primary challenges when using containerized workloads and the associated orchestrator for network simulation purposes is the presence of noise caused by management traffic on the network. Earlier versions of our simulator employed GreTap tunnels to isolate the simulation network from management traffic and other unrelated messages (Woltjen, et al., 2020). This was replaced by Virtual Extensible LAN (VXLAN)-based tunnels, as GreTap does not allow for redundant point-to-point connections between two network devices due to a lack

² Kubernetes: <https://kubernetes.io/>

³ Dockerfile reference docs: <https://docs.docker.com/engine/reference/builder/>

⁴ Docker Hub: <https://hub.docker.com>

of session identifiers. The use of VXLAN tunnels results in clean network captures only containing the traffic associated with the simulation.

3.2 Repeatable Simulation Scenarios

Our simulator allows the execution of fully repeatable simulation scenarios. Repeatable simulation scenarios are an important factor in verifying component behavior and in artificial intelligence (AI) training. This is made possible through the use of simulation sequences and shareable topologies as well as immutable component images. The Sequence defines a timed sequence of simulation steps, which can easily be repeated in every iteration of the simulation. A step can contain various commands, for example for starting a network-based attack. The Topology allows the sharing of a network environment with statically defined components and configurations. Immutable component images ensure that each simulation Node is at a fixed, known state at the beginning of the simulation. The combination of these three aspects ensure that a simulation run is not only repeatable, but also mostly reproducible. Currently, the following non-reproducible aspects are known:

- The system time gets passed from the host system to the Nodes by the container runtime, and subsequently differs between runs.
- Any naturally persistent system, like integrating the host or bridging external networks, cannot be controlled and reset by our simulator.
- The computational resources of the host system might influence the time needed for certain operations.
- The kernel of the host system, e.g. when running on a different machine, can cause slight differences.

3.3 Co-Simulation

One of the primary use-cases for our simulator rettij is to provide a network simulation for co-simulation scenarios managed by mosaik⁵. Our simulator provides an interface for control and data exchange through mosaik, as shown in Figure 2.

For example, this enables a scenario where our simulator and the power system modeling tool pandapower⁶ are co-simulating both the ICT and the physical energy component state. Within this scenario, control commands are passed to our simulator through the mosaik data exchange interface, transmitted over the network via a domain-specific communication protocol, and then passed on to pandapower via mosaik. This environment can be used to study the possibilities for and the physical impact of ICT-based attacks like FDIA and denial-of-service (DoS) on energy infrastructure.

In order to facilitate the control and data exchange through mosaik, our simulator implements a step mechanism that accepts input and runs timed commands when triggered. It also defines data retrieval logic. The mechanism is depicted in Figure 3. First, mosaik calls our simulator at a scheduled time, passing both that time and the registered inputs to the co-simulation interface of the simulator.

The received data is refined and passed on to the internal step logic of our simulator. This logic first writes the input data to the corresponding Nodes. Afterwards, the list of scheduled steps is queried for a step at the current co-simulation time. If a step is present, its commands will be executed. Once the execution is finished, our simulator will return its next scheduled time to mosaik. Since the network simulation is continuously running and not discrete, the next scheduled time is simply the smallest available increment of mosaik time units from the last call time (i.e. call time + 1). At the end of each time slot, mosaik will query for a set of data points from its simulators, depending on how they are connected in the scenario. Our simulator retrieves the corresponding values from its Nodes and returns them to mosaik.

⁵ Mosaik: <https://mosaik.offis.de/>

⁶ Pandapower: <https://www.pandapower.org/>

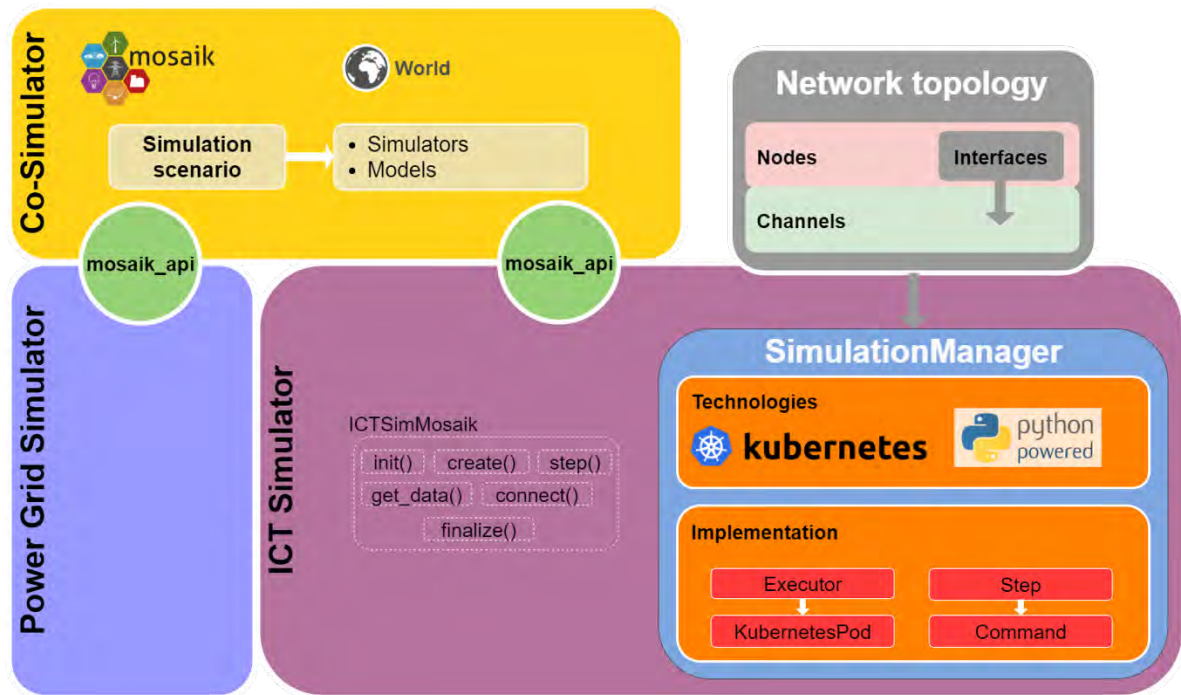


Figure 2: Overview of the interaction between our simulator and the co-simulator

One issue for our simulator that arises from its role as a “courier” in the co-simulation is the lack of information about the destination device for the data received from mosaik. Usually, mosaik connects one entity to another entity directly, and passes data via its data exchange interface. This is not possible with our simulator, as data from one simulator entity destined for another entity is not passed directly, but through the simulation network. To remedy this issue, our simulator provides a connect method similar to that of mosaik. This method allows the user to define (from within the mosaik scenario) where a message should be sent when a specific data point is updated. To facilitate the highest amount of flexibility, this functionality is constructed as a hook, meaning each component has the ability to specify custom logic as to how the information on data points and destinations is used within the simulation.

3.4 Distributed Simulation

One of the distinguishing features of our simulator is the capability to run simulations on distributed computing resources rather than a single device. This has two advantages: scaling, and integration of geographically distributed hardware.

While the simulation capacities on a single device are either fairly limited or come at very high purchasing costs, our simulator allows scaling the simulation over a cluster of various, inexpensive computing nodes. This cluster may be geographically distributed. Hardware labs containing physical systems in one location can be connected to hardware or computing resources in another location through our simulator. Both scaling and geographical distribution are enabled through the use of Kubernetes. The Container Networking Interface (CNI)⁷ allows for the creation of overlay networks over various host systems, resulting in a common network for all simulated components regardless of physical location of the underlying node. Depending on the requirements regarding distribution and the used cluster, an appropriate CNI can be used. As Kubernetes is meant for the orchestration of containerized workloads in large clusters, scaling and automated distribution of workloads across the cluster nodes are basic features.

⁷ Kubernetes Networking Plugins: <https://kubernetes.io/docs/concepts/extend-kubernetes/compute-storage-net/network-plugins/>

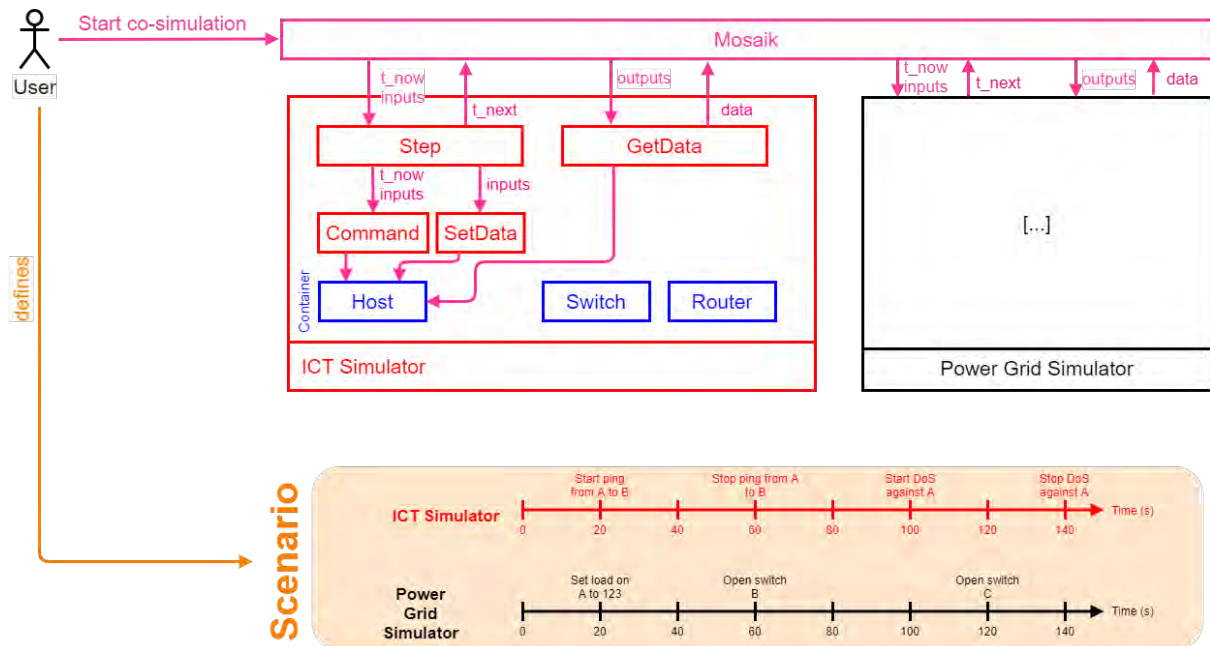


Figure 3: Co-simulation step mechanism

3.5 External Persistent Systems

In addition to container-based simulation Nodes, the systems running the Kubernetes cluster can also be attached to the simulation network via the topology. The host system can be connected via a layer 3 virtual IP interface, or act as a layer 2 bridge between the simulation network and its physical NICs. This allows for the integration of external, persistent networks and systems into the simulation. For example, a physical remote terminal unit (RTU) can be connected to a physical host in the same room, running a grid control center software, via a simulated WAN, eliminating the need for additional hardware. In addition to external persistent systems, Kubernetes-orchestrated virtual machines (VMs) are currently planned as an available node type as well.

4. Use Cases in Smart Grid Research Projects

Several research projects drive the development of our simulator through their requirements. All projects have a high demand to simulate ICT network infrastructures together with power grids. The mosaik co-simulator is used to combine and coordinate the different simulators, with our simulator being one of those simulators. Three of these project-related use cases for our simulator are outlined below.

4.1 Multi-Domain Attack Simulation with Adversarial Resilience Learning

A multi-domain analysis is performed in the project “Polymorphic agents as a cross-sectional software technology for analyzing the operational reliability of cyber-physical systems” (PYRATE). The goal is to analyze the impact of attacks in the domain’s *energy market*, *power grid* and *ICT* (Veith, et al., 2020). In order to systematically analyze how attacks within the three domains can influence each other, different scenarios are run in a co-simulation. One aspect analyzed will be the question of how an attacker can (legally) obtain the highest possible profit on the energy market. To achieve this, an extensive co-simulation is implemented, providing a local energy market, an electricity grid and the underlying communication network. The complete software stack also contains an AI attacker agent based on Adversarial Resilience Learning (ARL), which is used to attack the simulated domains under conditions that are as realistic as possible (Veith, et al., 2020). The attacker has access to sensors and actuators in all domains and is able to read and control them. In order to simulate the consequences of realistic ICT attacks, our simulator provides actuators and sensors to be used by the attacker. Actuators can manipulate data in network traffic or execute an attack such as temporary DoS attacks through a shutdown or reboot of a ICT device during the simulation. The provided sensors allow attackers to read data such as device states and data values that are sent through our network.

For this purpose, the network simulation needs a native co-simulation interface for mosaik and must allow simple integration of domain-specific components in order to provide sensors and actuators. Due to the AI

training for the attacker and the number of components, a large simulation is expected, requiring high performance of the simulated ICT devices at low costs.

4.2 Network Data Manipulation in Co-Simulation

The increasing integration of ICT in power systems is also part of the research project “Methods for grid operators to prevent, detect and respond to IT attacks and failures” (MEDIT), (van der Velde, et al., 2020). The projects aims to develop methods for actors in the electric power system to prevent, detect and react to ICT attacks and resulting failures.

In this context, our simulator is part of the research and validation environment to investigate complex interactions between ICT and power systems. Our simulator will be used as part of a co-simulation environment, and will also be integrated into a real power system laboratory. It therefore needs to be able to integrate vRTUs and real hardware RTUs controlling the power flow in the laboratory. In order to develop methods for detecting attacks on the power grid, the simulated ICT devices must behave as close to reality as possible. We therefore need to be able to carry out ICT attacks in realistic environments.

Multiple attack scenarios, such as man-in-the-middle (MITM) attacks on IEC60870-5-104 packets within the network simulation, will be performed in this project. Our simulator receives IEC 60870-5-104 packages from a power simulation and passes it to the corresponding vRTU within our simulator. The vRTU then sends the IEC 60870-5-104 data frames over its simulated network to their destination. On this basis, methods to detect attacks (e.g. intrusion detection system (IDS), monitoring etc.) can be evaluated in a holistic simulation. With this, common ICT attack detection and monitoring systems can be adapted to fit into smart grids, and new attack detections mechanisms can be evaluated for their applicability in the power grid domain.

Our simulator enables such these analysis and evaluations due to its co-simulation interface and the usage of containers that allow the integration of IDS and other systems.

4.3 Blockchain Simulation

Within the research project “Trusted blockchains for the open, smart energy grid of the future” (tbiEnergy), a local energy market is to be developed and implemented with the help of blockchain technology and the usage of additional hardware security. We find the currently available network simulators to be rather difficult to use in this context due to the large number of proprietary blockchain-specific components, which have to be simulated or implemented in the simulator specific language. Our simulator offers a simple integration of components into simulations without the need for extensive adjustment of the original components. In the context of this project, this allows testing the developed blockchain concept before the roll-out on real hardware. Because of the small-scale nature of the systems in the smart grid, and the complexity of interactions between them, hardware requirements for simulation systems are high. With its ability to run simulations distributed on a cluster, our simulator offers an advantage over classic simulators that can only use the computing resources of a single hardware system. Such a cluster could be scaled according to the specific simulation. This scalability in connection with the use of our simulator in co-simulations offers a solution for the simulation of complex scenarios within the smart grid. New approaches and concepts can thus be tested as in a realistic environment, even before the rollout on actual hardware. In the context of this project, our simulator thus offers the possibility to validate the countermeasures, which mitigate threats identified in an initial threat analysis, and to uncover any remaining weaknesses within the system.

5. Practical Application Case Study

In order to verify our simulator’s functionalities in a co-simulation environment, a practical application example was created. This example is meant to show the co-simulation with mosaik as well as our simulator’s networking capabilities.

For this purpose, a minimal scenario featuring a MITM-based message modification attack was implemented. This co-simulation scenario contains our simulator and two simple simulators, one for creating and sending and one for receiving and storing random data values (see Figure 4). The network topology for our simulator consist of a Node each for the sender and receiver simulators which contains a simple TCP sender/receiver program, as well as the network components required for three virtual LANs with routing between them. Additionally, an attacker component is connected to one of the switches. It is used to intercept and modify the traffic from the sender to the receiver Node.

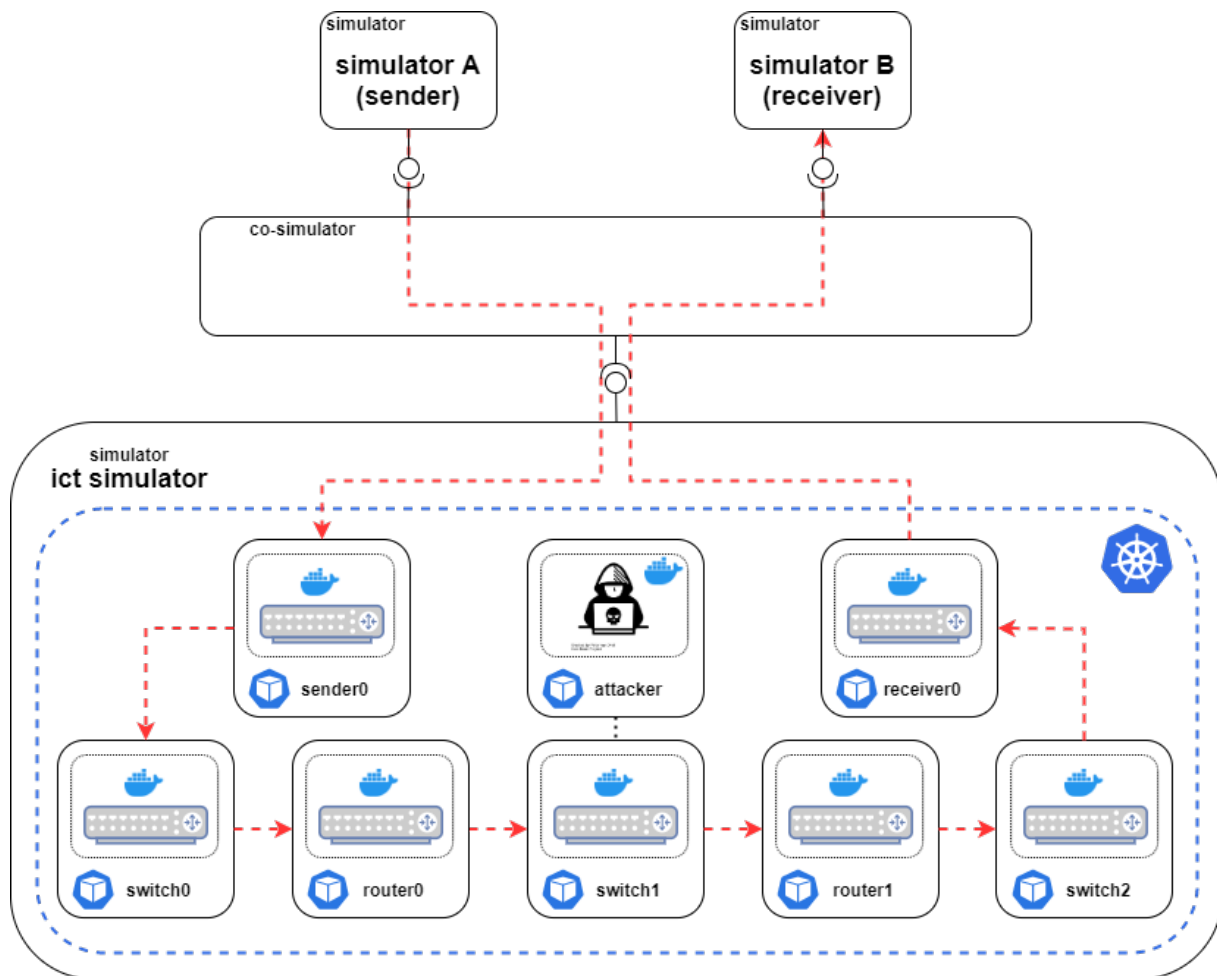


Figure 4: Dataflow co-simulation

During the simulation run, the sender simulator will continuously create a semirandom, incrementing integer value which is passed on the sender0 Node in our simulator rettij via mosaik. The TCP sender program on the Node then sends the data to the receiver0 Node via the network. The packet is then captured and modified by the attacker Node as part of the MITM-based message modification attack (similar to an FDIA). Once the packet reaches its destination on the receiver0 Node, it will be passed on to the receiver simulator where the data is then stored. In order to establish a baseline for the communication between the sender0 and receiver0 Node, the simulation sequence is defined in such a way that it starts a packet capture on switch1 at the very beginning of the simulation ($t=0$), and only triggers the attack at $t=2$. This way, unmodified traffic is captured for the first 2 time increments of the simulation.

```
Entity Model_sender_0 sent '4' for attribute 'value' at time 0.
Entity Model_receiver_0 received '4' for attribute 'value' at time 0.
Entity Model_sender_0 sent '6' for attribute 'value' at time 1.
Entity Model_receiver_0 received '6' for attribute 'value' at time 1.
Entity Model_sender_0 sent '16' for attribute 'value' at time 2.
Entity Model_receiver_0 received '16' for attribute 'value' at time 2.
Entity Model_sender_0 sent '22' for attribute 'value' at time 3.
Entity Model_receiver_0 received '-22' for attribute 'value' at time 3.
Entity Model_sender_0 sent '28' for attribute 'value' at time 4.
Entity Model_receiver_0 received '-28' for attribute 'value' at time 4.
Entity Model_sender_0 sent '37' for attribute 'value' at time 5.
Entity Model_receiver_0 received '-37' for attribute 'value' at time 5.
```

Figure 5: Data send and received by the simulator

The attack was run for 12 time increments, during which the attacker simply inverted the intercepted values ($\text{value} * -1$) and then forwarded the packets to their original destination. A section of the data sent and received

by the sender and receiver simulators is shown in Figure 5. There, the beginning of the message modification attack can be seen between t=2 and t=3.

```

10:07:28.541382 02:00:00:00:ff:01 > 02:00:00:00:ff:02, ethertype IPv4 (0x0800),
length 80: 192.168.0.10.44984 > 192.168.1.10.9999: Flags [P.], seq
785500028:785500042, ack 695056372, win 510, options [nop,nop,TS val
3366767962 ecr 1158610599], length 14
0x0000: 4500 0042 63a7 4000 3f06 55aa c0a8 000a E..Bc.@.?..U....
0x0010: c0a8 010a afb8 270f 2ed1 c77c 296d b7f4 .....'....|)m..
0x0020: 8018 01fe 8299 0000 0101 080a c8ac cd5a .....Z
0x0030: 450e fea7 7b22 7661 6c75 6522 3a20 3136 E...{"value":.16
0x0040: 7d0a                                     }.

10:07:36.946983 02:00:00:00:ff:01 > 02:00:00:00:ff:05, ethertype IPv4 (0x0800),
length 80: 192.168.0.10.45090 > 192.168.1.10.9999: Flags [P.], seq
1836877175:1836877189, ack 1944996579, win 510, options [nop,nop,TS val
3366776367 ecr 1158619003], length 14
0x0000: 4500 0042 e4a8 4000 3f06 d4a8 c0a8 000a E..B..@.?.....
0x0010: c0a8 010a b022 270f 6d7c 8577 73ee 4ae3 .....".'.m|.ws.J.
0x0020: 8018 01fe 8299 0000 0101 080a c8ac ee2f ...../
0x0030: 450f 1f7b 7b22 7661 6c75 6522 3a20 3232 E...{"value":.22
0x0040: 7d0a                                     }.

10:07:36.949176 02:00:00:00:ff:05 > 02:00:00:00:ff:02, ethertype IPv4 (0x0800),
length 80: 192.168.0.10.45090 > 192.168.1.10.9999: Flags [P.], seq 0:14, ack
1, win 510, options [nop,nop,TS val 3366776367 ecr 1158619003], length 14
0x0000: 4500 0042 e4a8 4000 3e06 d5a8 c0a8 000a E..B..@.>.....
0x0010: c0a8 010a b022 270f 6d7c 8577 73ee 4ae3 .....".'.m|.ws.J.
0x0020: 8018 01fe f0fa 0000 0101 080a c8ac ee2f ...../
0x0030: 450f 1f7b 7b22 7661 6c75 6522 3a20 2d32 E...{"value":.-2
0x0040: 327d                                     2}

```

Figure 6: Packet capture on switch 1

On the network level, this is also visible in the output of tcpdump on switch1, as shown in Figure 6. While the transmitted value 16 at t=2 is transmitted between router0 (02:00:00:00:ff:01) and router1 (02:00:00:00:ff:02) with no interception, value 22 at t=3 is first passed on to the attacker (02:00:00:00:ff:05) where it is modified and then sent to its original destination via router1.

Using this practical application example, we were able to show that our simulator rettij is able to exchange data with other simulators through the co-simulation interface, perform a real-world ICT attack on a realistic network topology utilizing both switches and routers as well as user-defined components and analyze the resulting traffic without any network noise. In terms of practical applicability, the demonstrated scenario is close to actual attacks on ICT infrastructure in the power grid, during which control commands and value measurements are modified or suppressed in order to influence the associated physical systems.

6. Conclusion

In this publication, the current challenges in the area of smart grid co-simulation of ICT with regard to its security were highlighted. In particular, the increasing attack surface through ICT must be considered in future simulations in order to gain new insights and to be able to adapt and develop new ICT security tools to the needs of the smart grid. This requires observations spanning multiple domains, which in turn requires an ICT simulator that can be used in co-simulations. Consequently, our network simulator was presented, providing a co-simulation interface for mosaik. The network simulator makes use of containers and Kubernetes which allows a simple integration of various software components, such as vRTUs, as well as distributed ICT simulation. Use cases from various research projects are shown to demonstrate the need as well as the benefits of a distributed, co-simulation capable ICT simulator. The introduction of real-world ICT attacks into a multi-domain smart grid co-simulation was demonstrated by a practical example, presenting some of the distinguishing features of our network simulator.

References

- Doleski, O. D., 2017. *Wie sich die Energiewirtschaft im Zeitalter der Digitalisierung verändert*. Wiesbaden: Springer Fachmedien Wiesbaden GmbH 2017.
- Yan, Y., Qian, Y. & Sharif, H., 2013. A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges. Band 20, pp. 5-20.

- Tuinema, B. W. et al., 2020. *Cyber-physical system modeling for assessment and enhancement of power grid cyber security, resilience, and reliability*. In *Probabilistic Reliability Analysis of Power Systems*. s.l.:Springer.
- Soares, J. et al., 2008. *A Cosimulation Architecture for Power System, Communication, and Market in the Smart Grid*, s.l.: s.n.
- Moulema, P., Yu, W., Griffith, D. & Golmie, N., 2015. On Effectiveness of Smart Grid Applications Using Co-Simulation. In: *2015 24th International Conference on Computer Communication and Networks (ICCCN)*. s.l.:IEEE, pp. 1-8.
- Veith, E. M. et al., 2020. *The Adversarial Resilience Learning Architecture for AI-based Modelling, Exploration, and Operation of Complex Cyber-Physical Systems*. online, IARIA.
- Woltjen, T., Gritzan, G., Kathmann, P. & Sethmann, R., 2020. *Simulationsumgebung für IKT-Netze zur Cyber-Abwehr: IKT-Netzsimulation zur Prävention von Angriffen auf das Energienetz*. Berlin, VDE VERLAG GMBH, pp. 233-239.
- van der Velde, D. et al., 2020. *Methods for Actors in the Electric Power System to Prevent, Detect and React to ICT Attacks and Failures*. Gammarth, Tunisia, IEEE.
- M. Igure, V., A. Laughter, S. & D. Williams, R., 2006. *Security issues in SCADA networks*. 7 Hrsg. s.l.:Computers & Security.
- I. T. F. o. I. T. et al., 2019. Interfacing Power System and ICT Simulators: Challenges, State-of-the-Art, and Case Studies. *IEEE Transactions on Smart Grid*, pp. 14-24.

Towards Detection of Selfish Mining Using Machine Learning

Matthew Peterson, Todd Andel and Ryan Benton

The University of South Alabama, Mobile, USA

mp1524@jagmail.southalabama.edu

tandel@southalabama.edu

rbenton@southalabama.edu

Abstract: Selfish mining is an attack against a blockchain where miners hide newly discovered blocks instead of publishing them to the rest of the network. The selfish miners continue to mine on their private chain while the honest miners waste resources mining on a shorter chain. According to the blockchain protocol, a longer chain takes precedent and shorter chains are discarded which allows the selfish miners to gain an advantage by keeping their chain secret. This attack can be used by malicious miners to earn a disproportionate share of the mining rewards or in conjunction with other attacks to steal money from cryptocurrency exchanges. Several of these attacks were launched in 2018 and 2019 with the attackers stealing as much as \$18 Million. Developers made several different attempts to fix this issue, but the effectiveness of the fixes is currently unknown. Although this attack is possible against both Proof-of-Work and Proof-of-Stake blockchains, this research concentrates on detection in Proof-of-Work blockchains. As is difficult to evaluate security advances in the real-time blockchain, it is imperative to focus on simulation to evaluate blockchain security properties. To this end, we extend a blockchain simulator and add the ability to simulate selfish mining attacks. Several existing simulators are examined before choosing SimBlock for this research. Our goal is to identify the factors that identify selfish mining. Using existing research, we choose several factors that could identify an attack in an unlaunched state, an active state, or historically. We plan to use simulated data to train a machine learning model to detect selfish mining. Using the modified simulator, we generate training and test data for unlaunched and active attacks. For historical attacks, we will use historical data from known selfish mining attacks. While some existing research has examined the detection of selfish mining, it only examines active attacks. In this paper, we seek to lay the groundwork for future research into detecting attacks that are unlaunched, active, or historical.

Keywords: blockchain, selfish mining, proof-of-work, detection, simulation, machine learning

1. Introduction

Blockchains, at their core, are data structures that contain cryptographically-linked transactions. This data structure enables a peer-to-peer network to reach consensus about the state of transactions without requiring a central authority. Blockchain's first and most compelling application is that of digital currency. Although the beginnings of a blockchain-based digital currency already existed in the form of a white paper, the idea gained momentum when Bitcoin first appeared (Dai, no date). In order to make cryptocurrencies work, a mechanism is needed that enables network participants to reach consensus about the state of the blockchain. Each network participant keeps a copy of the ledger of transactions and without a consensus mechanism, one could compromise its integrity. Blockchains use many different consensus mechanisms, but the two most common ones are Proof-of-Work (PoW) and Proof-of-Stake (PoS).

In 2013 authors Eyal and Sirer discovered a vulnerability in the Proof-of-Work consensus algorithm that allowed malicious users to launch an attack against Bitcoin (Eyal and Sirer, 2018). Dubbed the selfish mining attack, this attack allows a miner, or a pool of miners, to earn more than their fair share of the mining rewards by withholding discovered blocks from the rest of the network. The attacker will publish these hidden blocks later to undo all the work of the other miners. Although other consensus algorithms are also vulnerable to selfish mining, this research focuses on the PoW algorithm. Selfish mining broke onto the national scene when it was used with double-spend attacks to steal cryptocurrency from cryptocurrency exchanges. In a double-spend attack, the attackers seek to spend their cryptocurrency twice by manipulating the protocol. The attackers will create two transactions where one transaction sells the cryptocurrency to an exchange while the second sends the same currency to an attacker-owned wallet.

These attacks started happening in 2018 and continued into 2019. In April 2018, the Verge cryptocurrency was attacked twice with the attacker getting away with as much as \$2.8 Million (*A History of 51 Percent Attacks on Blockchains and Cryptocurrencies*, 2019). That same month, someone attacked Monacoin and stole another \$90,000. The very next month the biggest attack to date happened to Bitcoin Gold when attackers used selfish mining and double-spending to steal \$18 Million from cryptocurrency exchanges (Hertig, 2018). Following the Bitcoin Gold attack, Litecoin Cash was attacked in the same month, Zencash was attacked in June, and lastly,

Ethereum Classic was attacked in January of 2019 (Nesbitt, 2019; *A History of 51 Percent Attacks on Blockchains and Cryptocurrencies*, 2019).

Despite the growth of selfish mining attacks, there is limited research into their detection. Several authors have suggested ways to detect this attack but stopped short of doing any research into it (Göbel *et al.*, 2016; Eyal and Sirer, 2018). A few papers researched detection but only focused on a single factor or only worked on detection in a single stage (Chicarino *et al.*, 2020; Kędziora *et al.*, 2020; Liu *et al.*, 2020). Blockchain developers have released patches to prevent selfish mining, but without reliable detection, it is impossible to tell if these patches were successful. Detection of selfish mining is needed in the historical blockchain data as well as in the live network. Our goal is to examine all the factors that indicate a selfish mining attack and attempt to identify the significant ones by using machine learning. Selfish mining attacks exist in three distinct stages: unlaunched, active, and historical. Detection is the first and most important step in preventing an attack and we attempt to see if it is possible to detect an attack in all three of these stages. This paper lays the groundwork for further research into this topic.

In summary, our goals are as follows:

1. Determine the significant factors that identify an unlaunched selfish mining attack
2. Determine the significant factors that identify an active selfish mining attack
3. Determine the significant factors that identify a historical selfish mining attack from existing blockchain data

2. Literature Review

2.1 Proof-of-Work (PoW) Consensus Algorithm

To understand how selfish mining works, it is important to understand how mining works. Miners use a process called Proof-of-Work (PoW) to mine new blocks. Originally designed by Adam Back to counter denial-of-service attacks, PoW allows a peer-to-peer network to agree on the state of a blockchain (Back, 2002). At its core, a blockchain is a data structure of cryptographically-linked transactions. When someone wants to make a purchase using Bitcoin, he broadcasts the transaction to the network. Transactions are picked up by miners who bundle transactions together into blocks. Miners create blocks by taking the transactions and running them through a hashing algorithm until they find a hash that matches the network-defined target value.

This process of finding a valid hash is very computationally expensive and time-consuming. Miners are incentivized to create new blocks by receiving newly minted cryptocurrency when their block is accepted by the network.

The blockchain can fork in two when two miners discover a block at the same time. For example, if Miner A and Miner B both find a block and broadcast them to the network, Miner C will accept the block he hears about first. Blockchains resolve this issue with the principle of the longest chain.

The longest chain principle causes the network participants to adopt the longest available blockchain (the one with the most Proof-of-Work). If a miner is mining on a chain with a height of n and is notified about a valid chain with a height of $n + 1$, he will discard his current chain and switch to the new chain. This means the fork is resolved by whichever version of the blockchain finds the next block first (Nakamoto, 2008).

2.2 Selfish Mining

Selfish mining manipulates the way consensus algorithms work to gain a disproportionate share of the mining rewards. Under normal operation, miners work to discover a new block to add to the blockchain. Once found, this new block is broadcast to the rest of the network and each node adds it to their local blockchain. Selfish mining occurs when a miner keeps a discovered block hidden instead of letting the rest of the network know. On the surface, this seems counterintuitive. If the miner refuses to broadcast the block, they run the risk of another miner finding a block and rendering their work useless. If, however, a selfish miner discovers two or more blocks before the rest of the network, they cause the rest of the network to waste their time mining on a shorter chain. The minute the network gets close to the selfish miner's lead, they will release the longer chain on the network and reverse all the blocks mined by the honest miners using the longest chain principle discussed in the previous section.

The selfish mining attack was first discovered by Eyal and Sirer who showed that miners could attack a blockchain by keeping their mined blocks hidden. As shown in Table 1, Eyal and Sirer defined the parameter α as the percentage of the total hash rate controlled by the selfish miners and parameter γ as the percentage of the honest nodes who adopt the selfish miner's block in the case of a network fork. Their papers showed that miners with $\alpha \leq 25\%$ and $\gamma = 50\%$ have enough monetary incentive to launch a selfish mining attack (Eyal and Sirer, 2018). Since this discovery, there have been multiple papers that examine different ways to fix selfish mining, extend the original selfish mining model, or dispute the proposal that selfish mining is incentive compatible.

2.3 Detecting Selfish Mining

Little research has been done in detecting selfish mining. In their original paper, Eyal and Sirer noted the difficulties facing accurate detection. They claimed that the two biggest indicators of selfish mining are the orphan block rate and the timings between successive blocks (Eyal and Sirer, 2018).

The orphan block rate is difficult to measure as the blockchain protocol discards orphaned blocks, so unless the orphaned blocks are captured in real-time, the data is lost forever. Another factor that makes detection difficult is the naturally occurring orphan block rate. Forks naturally occur in the blockchain and the protocol is designed to converge back to a single chain by orphaning one of the two chains. The difficulty is knowing when the orphaned blocks are part of an attack vs when they occurred naturally.

The natural orphan block rate is relatively low. Göbel, Krzesinski, and Taylor examined the orphan block rate in the absence of network delay and found that it hovers around 0.53 blocks per day in simulations and 2-3 blocks on the real Bitcoin network (Göbel *et al.*, 2016). They also examined the orphan block rate as it relates to selfish mining by simulating 70 days of blockchain activity for 1,000 nodes. They found that an increase in the number of blockchain forks suggested someone was engaged in selfish mining.

While this research showed several good selfish mining indicators, some factors make it difficult to implement with live network traffic. The only nodes who know of a blockchain fork are the nodes on the edge of the fork. If a node receives a blockchain from a peer that is the same length as his current blockchain, he will discard it. Similarly, if he sends his blockchain to the same peer, that peer will discard the blockchain. This creates a natural dividing line through the peer-to-peer network where only the nodes on the line will know about the fork. For a node to discover a spike in the fork rate, it would have to be connected to every other node on the network or deploy multiple listener nodes.

Decker and Wattenhofer also examined the rate at which the Bitcoin network forked and thereby produced orphan blocks. After collecting 10,000 blocks, they found that the blockchain forked 169 times for a rate of 1.69% (Decker and Wattenhofer, 2013). This averages out to 2.4 orphan blocks per day which correlate with the findings by Göbel, Krzesinski, and Taylor (Göbel *et al.*, 2016). This number assumes that only one block was orphaned during the fork but multiple blocks can get orphaned at once. Additionally, they found that larger block sizes made the network more susceptible to forking due to the time spent validating the larger blocks.

The second indicator of selfish mining is the timings between successive blocks. If two or more blocks are discovered in quick succession, someone is likely to be engaged in selfish mining. The blocks in a PoW blockchain are exponentially distributed so the sudden deviation of finding two blocks close in time to each other is significant. While unlikely, it is worth noting that finding two blocks in quick succession is possible without a selfish mining attack.

One of the only papers to look at the block distribution was an unpublished paper by Fangyang Cui. This researcher graphed the distribution of Bitcoin blocks and found that the Bitcoin network follows the expected exponential distribution (Cui, 2015). This research is confirmed by another article that used the blockchain.info API. This author did not find any evidence of selfish mining but found that the blocks matched the expected exponential distribution.

The third indicator of selfish mining is the miner's average revenue-per-hour. Göbel, Krzesinski, and Taylor suggested that a drop in the revenue-per-hour of miners could indicate selfish mining but stopped short of proving their assumptions (Göbel *et al.*, 2016).

There are a few papers that examine how to detect selfish mining. The first, written by Chicarino, Albuquerque, Jesus, and Rocha, looked at detecting selfish mining using the block height of forks (Chicarino *et al.*, 2020). Using a simulator, they modeled a selfish mining attack and captured the output data. By examining the block forks, they determined that it was possible to detect an active blockchain attack based on the height of the forked chain. If the forked chain's block height was greater than or equal to two, they classified it as an attack. The authors conclude by stating that research is needed into reducing false positives from honest chains with a height greater than two.

Another paper examined different combinations of parameters and how they influenced the vulnerability of the blockchain network (Kędziora *et al.*, 2020). The parameters under study were the difference between the selfish miner's chain and the honest chain, the number of connections made by the selfish miners, and the total hashing power of the selfish miners. They examined how the choice of different parameters affected the distribution of block extraction between selfish and honest miners.

Lastly, Liu *et al.* examined the theoretical state space of a selfish mining attack by extending a Markov Chain simulation (Liu *et al.*, 2020). Like others, their research looked at an active attack and found that the number of attack states stayed mostly consistent regardless of the attacker's mining power.

3. Method

Our goal is to examine all the factors that identify a selfish mining attack and find significant ones. As we are unsure of what factors will prove significant, we plan on using a Random Forest machine learning algorithm and training it using simulated blockchain data. We view the detection of selfish mining as a classification problem (benign or selfish) and a Random Forest allows us to classify the data and see the feature importance. Using this simulated data, we plan on seeing if the significant factors can detect an attack in all three stages.

3.1 Simulating blockchain data

To train the model, we plan on simulating 70 days' of blockchain data by using SimBlock. SimBlock is a peer-reviewed blockchain simulator written in Java (Aoki *et al.*, 2019). To generate the correct data, SimBlock needed modification to simulate selfish mining. We made several modifications to the program to add selfish mining nodes to the simulated network a copy is available on our GitHub upon request ([themattman18/simblock](#), no date). The modifications to SimBlock followed the guidelines from the original selfish mining paper by Eyal and Sirer for the selfish miner's hash rate, and block propagation.

Table 1: Table of Notations

Symbol	Description
s	The selfish miners who are attacking the network
h	The honest miners who are following the Bitcoin mining protocol
α	The percentage of the total network hash rate owned by the selfish miners
γ	The percentage of honest nodes who adopt the selfish miner's block in the event of a network fork

The newly added selfish mining node follows Eyal and Sirer's original selfish mine algorithm. The algorithm defines the following states for selfish miners (s) and honest miners (h):

- s gains a one-block lead $\rightarrow s$ keeps block secret
- s has a one-block lead but h finds block $\rightarrow s$ releases secret block, causing a fork
- s has no lead and h finds block $\rightarrow s$ adopts h block
- s has a two-block lead and h finds block $\rightarrow s$ releases their longer chain onto the network
- s has a lead $>$ two blocks and h finds a new block $\rightarrow s$ does nothing

If s has a one-block lead and h discovers a new block, s will release their block onto the network to intentionally create a fork. The miners who hear about s block first will adopt it while those who hear about h block first will adopt that one. Selfish mining becomes profitable when the attackers can get $\frac{1}{2}$ of the network to adopt their block in this situation (Eyal and Sirer, 2018). The number of nodes who adopt s block is represented by the symbol γ and the nodes who adopt h block is $(1 - \gamma)$.

To launch a successful selfish mining attack, the selfish miners need to get $\frac{1}{2}$ of the network to adopt their block in the case of a blockchain fork. The SimBlock program connects nodes based on the observed distribution of connections in Bitcoin in the year 2015 taken from previous research (Gervais *et al.*, 2016; Miller *et al.*, no date).

If a simulation node is randomly selected into a region where the nodes were well connected, the simulation will connect that node to a larger number of other nodes. The better connected a selfish mining node is, the higher γ will end up being.

The total hash rate of the selfish miners is indicated by α with honest nodes controlling $(1 - \alpha)$ of the hash rate. As discussed previously, if the blockchain forks due to competing blocks between the selfish and honest miners, the selfish miners become more profitable as a higher percentage of honest nodes accept and propagate their block. If $\gamma = 0.5$ ($\frac{1}{2}$ of the honest nodes accept the selfish miner's block) then the selfish miners only need 25% of the network hash rate to successfully launch a selfish mining attack ($\alpha = 0.25$). We modified SimBlock and gave the newly created selfish mining node 25% of the total network mining power. Although a selfish mining attack can succeed between the ranges $0.5 > \alpha > 0$ we went with the threshold of $\alpha = 0.25$ as defined by Eyal and Sirer (Eyal and Sirer, 2018).

3.2 Generation of Training and Test Data

To generate the training data, we simulated 70 days' worth of blockchain data for 600 network nodes. Seventy days of data is the generally accepted amount of time for gathering data.

Our test data comes from two different places. The first set of test data comes from the modified SimBlock network simulator. We generated 30 days of data and categorized it into benign and selfish and plan on using it to verify the accuracy of the model. The second set of data, used to test the ability to detect historical attacks, will come from the Bitcoin Gold blockchain. Since blockchain data is not stored in a human-readable format, a program was needed to format the data into the input needed by the machine learning model. Bitcoin Gold hosts a block explorer that allows anyone to browse the latest released block on the network (*Bitcoin Gold Block Explorer*, no date). This website presents a way to view the latest blockchain data and allows users to jump to a date to view the recorded blocks on that day.

Two separate periods were gathered from the block explorer website. The first benign set of data was gathered from 5/15/2018 approximately three days before the first known selfish mining attack. 155 blocks were gathered from block height 529040 to 529194. The second set of data came from a selfish mining attack on 5/18/2018-5/19/2018. We gathered a total of 25 blocks ranging from block height 529022 to 529047.

3.3 Feature Selection

The first feature we selected is the block height. As shown by Chicarino, Albuquerque, Jesus, and Rocha, the block height of a fork can be used to detect a launched selfish mining attack (Chicarino *et al.*, 2020). If a blockchain fork had a block height greater than or equal to two, they classified it as an attack. Unfortunately, this resulted in some false negatives from forks that only had one block and it cannot detect historical or unlaunched selfish mining attacks.

The second feature is the timing between blocks. The discovery of new blocks is Poisson distributed which means that the time it takes to discover a new block is independent of when the last block was found. Although the act of finding a new block is an independent variable, the average time it takes is kept at 10 minutes per block. This exponential distribution makes it difficult to tell if two blocks released back-to-back were selfishly mined or just part of the normal mining process but is a useful feature.

Multiple papers have discussed the importance of the fork rate on the detection of selfish mining (Göbel *et al.*, 2016; Eyal and Sirer, 2018). Forks create a natural rift in the network where only the nodes on the edge of this rift know about the fork and detecting all forks would require connecting to all the network nodes. To bypass this limitation, we ignore this restriction in our simulation and report all forks.

The transaction count is another feature we selected. Decker and Wattenhofer measured the effects of block size on propagation speed and noticed a strong correlation between the two (Decker and Wattenhofer, 2013). Smaller blocks propagate through the network faster than larger ones and a selfish miner wants to propagate their blocks as fast as they can to ensure that most of the nodes build on top of theirs. Honest nodes, on the other hand, want to include as many transactions as possible to gain more money from transaction fees. This could result in a large disparity between the transaction counts in honestly mined blocks and the selfishly mined blocks.

The next feature we selected is the transaction output wallet address. An active selfish mining attack could double-spend cryptocurrency and the attackers will send the double-spent money back to a wallet address that they control. While it is best practice to create a new wallet address for every new transaction, it is not required. A selfish miner could continue to selfishly mine and send transactions to the same wallet address.

Another feature we selected is the coinbase transaction which sends the reward for mining a new block to the miner-specified wallet address. Regardless of double-spent transactions, attackers want to receive the financial reward for mining a new block. Just like in the other transactions, the miner can supply a new wallet address for every new coinbase transaction. However, in some observed selfish mining attacks, the attacker repeatedly sent the coinbase transaction to the same wallet address (*Bitcoin Blocks solved On 2018-05-19 | Insight*, no date).

Göbel, Krzesinski, and Taylor observed the importance of the revenue-per-hour as an indicator and calculated it as (*block rewards / percentage of the network hashing power*) (Göbel *et al.*, 2016). When selfish mining increases the honest node's revenue-per-hour will drop. As implied by the name, using the revenue per hour is a slow detection process because it works on the hourly average. The total, real network mining power can fluctuate but the difficulty itself is only adjusted once every two weeks which makes it difficult to know how much mining power there is on a network at any given point.

Using these features, we train the machine learning model to see if it is possible to detect selfish mining. Future research will examine the effectiveness of these features.

4. Review

In this paper, we lay the groundwork for building a machine learning model that can be used to detect unreleased, active, and historical selfish mining attacks. We extended the SimBlock blockchain simulator by adding the ability to simulate selfish mining attacks. By using the updated SimBlock, we generated training and test data for selfish mining attacks which we plan on using to train our machine learning model. The test data for the historical attacks come from examples of actual selfish mining attacks that were launched against the Bitcoin Gold blockchain. The next step is to finishing training the machine learning model and measure its effectiveness.

References

- A History of 51 Percent Attacks on Blockchains and Cryptocurrencies* (2019) *Blocks Decoded*. Available at: <https://blocksdecoded.com/history-51-percent-attacks/> (Accessed: 26 June 2019).
- Aoki, Y. *et al.* (2019) 'SimBlock: A Blockchain Network Simulator', in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 325–329. doi:10.1109/INFCOMW.2019.8845253.
- Back, A. (2002) 'Hashcash - A Denial of Service Counter-Measure', p. 10.
- Bitcoin Blocks solved On 2018-05-19 | Insight* (no date). Available at: <https://explorer.bitcoingold.org/insight/blocks-date/2018-05-19> (Accessed: 19 April 2020).
- Bitcoin Gold Block Explorer* (no date). Available at: <https://explorer.bitcoingold.org/insight/> (Accessed: 19 April 2020).
- Chicarino, V. *et al.* (2020) 'On the detection of selfish mining and stalker attacks in blockchain networks', *Annals of Telecommunications*, 75(3–4), pp. 143–152. doi:10.1007/s12243-019-00746-2.
- Cui, F. (2015) 'Detecting Selfish Mining in Bitcoin', *unpublished*, p. 4.
- Dai, W. (no date) *b-money*. Available at: <http://www.weidai.com/bmoney.txt> (Accessed: 15 February 2020).
- Decker, C. and Wattenhofer, R. (2013) 'Information propagation in the Bitcoin network', in *IEEE P2P 2013 Proceedings. 2013 IEEE Thirteenth International Conference on Peer-to-Peer Computing (P2P)*, Trento, Italy: IEEE, pp. 1–10. doi:10.1109/P2P.2013.6688704.
- Eyal, I. and Sirer, E.G. (2018) 'Bitcoin Mining Is Vulnerable', *Communications of the ACM*, 61(7), p. 8.
- Gervais, A. *et al.* (2016) 'On the Security and Performance of Proof of Work Blockchains', in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. the 2016 ACM SIGSAC Conference*, Vienna, Austria: ACM Press, pp. 3–16. doi:10.1145/2976749.2978341.
- Göbel, J. *et al.* (2016) 'Bitcoin blockchain dynamics: The selfish-mine strategy in the presence of propagation delay', *Performance Evaluation*, 104, pp. 23–41. doi:10.1016/j.peva.2016.07.001.
- Hertig, A. (2018) *Blockchain's Once-Fearful 51% Attack Is Now Becoming Regular*, *CoinDesk*. Available at: <https://www.coindesk.com/blockchains-feared-51-attack-now-becoming-regular> (Accessed: 24 February 2019).
- Kędziora, M. *et al.* (2020) 'Analysis of Blockchain Selfish Mining Attacks', in Borzieski, L., Świątek, J., and Wilimowska, Z. (eds) *Information Systems Architecture and Technology: Proceedings of 40th Anniversary International Conference on Information Systems Architecture and Technology – ISAT 2019*. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 231–240. doi:10.1007/978-3-030-30440-9_22.

- Liu, Z. *et al.* (2020) 'A Security Detection Model for Selfish Mining Attack', in Zheng, Z. *et al.* (eds) *Blockchain and Trustworthy Systems*. Singapore: Springer Singapore (Communications in Computer and Information Science), pp. 185–195. doi:10.1007/978-981-15-2777-7_16.
- Miller, A. *et al.* (no date) 'Discovering Bitcoin's Public Topology and Influential Nodes', p. 17.
- Nakamoto, S. (2008) 'Bitcoin: A Peer-to-Peer Electronic Cash System', p. 9.
- Nesbitt, M. (2019) *Ethereum Classic (ETC) is currently being 51% attacked*, *The Coinbase Blog*. Available at: <https://blog.coinbase.com/ethereum-classic-etc-is-currently-being-51-attacked-33be13ce32de> (Accessed: 24 February 2019).
- themattman18/simblock* (no date) *GitHub*. Available at: <https://github.com/themattman18/simblock> (Accessed: 13 April 2020).

Specialised Media Monitoring Tool to Observe Situational Awareness

Heloise Pieterse¹, Carien Van 't Wout², Zubeida Khan¹ and Chris Serfontein¹

¹Council for Scientific and Industrial Research, Pretoria, South Africa

²Capgemini, Amsterdam, Netherlands

hpieterse@csir.co.za

carien.wout@gmail.com

zdawood@csir.co.za

cserfontein@csir.co.za

Abstract: The 21st century witnessed the rise of the digital age via the evolution of the Internet. A direct result of the established digital era is the growth of digital media. Although traditional media, such as print and broadcasting, remains relevant, the low accessibility and limited reach impact the effectiveness of such media. In comparison, digital media, such as social media, provides access to vast collections of open-source information. The open-source information available on digital media platforms is current and easily accessible. However, conversion of open-source information into intelligence becomes a time-consuming and quite ineffective process due to the large quantities of information available. Open-source intelligence (OSINT) tools attempt to address such limitations by offering technological solutions that provide access to up to date, as well as easily searchable digital information. Thus, OSINT tools, such as Shodan and Maltego, streamline the conversion of open-source information into intelligence. This paper introduces the design of a new but specialised media monitoring tool to assist with the monitoring of open-source information towards enabling situational awareness. The key design feature of this new media monitoring tool is to offer versatility by guiding the collection of open-source information. Controlling the amount of information gathered in turn improves the efficiency of the processes used to extract intelligence. The obtained intelligence, thus, enables the observation of situational awareness of chosen aspects within a secure environment. Also discussed in this paper is the motivation behind the design of the specialised media monitoring tool, as well as various design considerations, are taken to ensure effective usage of the tool. The paper concludes by discussing the various benefits offered by the specialised media monitoring tool concerning observing situational awareness.

Keywords: Open-Source Information, Digital Media, Social Media, Monitoring, Intelligence, Situational Awareness

1. Introduction

The digital age ushered in a new era characterised by a sudden change from traditional industrial-based processes to an economy primarily driven by information and communication technology (ICT) (Forsyth, 1999). The success of the digital age is due to the evolution of the Internet, which permits the retrieval, sharing and transmission of information via digital media (Haris, 2016). Digital media, such as online newspapers, discussion boards and social media, provides access to vast collections of open-source information. Open-source information is best defined as “unclassified data available to the public” (Williams & Blum, 2018) that anyone can lawfully obtain by request or observation. Lately, open-source information has become increasingly valued (Noble, 2004) due to the proliferation of such information on the Internet, which is current and easily accessible. However, due to the sheer volume of readily available open-source information, conversion of the information into intelligence can be laborious and ineffective.

Technological solutions, such as Open-Source Intelligence (OSINT) tools, attempt to address the challenge of obtaining vast quantities of open source information for intelligence. The purpose of OSINT tools is to assist with the automated retrieval, collection, filtering and distribution of open-source information. Media monitoring systems and real-time search engines, which are examples of OSINT tools, provide analysts with a vast influx of open-source information from multiple data media sources across the world (Best, 2011). However, the present challenge is no longer accessing timely information but processing and analysing that information to obtain intelligence. Thus, to avoid information overload (Best, 2011) and enable the collection of highly accurate information (disregarding fake information), it is necessary to guide the collection of open-source information. Although various OSINT tools currently exist, such tools either limit the collection of open-source information to a single platform or collect from multiple digital sources, which impacts the effective processing and analysis of information. Therefore, sophisticated intelligence tools are required that provide analysts with access to rapid but relevant information, as well as techniques to produce intelligence insight and attain insight.

This paper introduces the design of a new specialised media monitoring tool that allows for directed, as well as geographically positioned searches to be conducted. The purpose of the specialised media monitoring tool is to assist analysts to observe situational awareness, which is defined as having awareness about what is happening in an environment and using the information to make informed decisions (Gheisari, Irizarry, & Horn, 2010).

To overcome known limitations of existing OSINT tools, the specialized media monitoring tool incorporates four distinct features: (i) monitoring pre-selected digital media platforms, (ii) geographically guide searches for information, (iii) direct the extraction of information, and (iv) structuring of the collected information. By guiding the collection of open-source information, the amount of information gathered is controlled, which in turn improves the efficiency of the deployed data mining and pre-analysis techniques to extract intelligence. The intelligence obtained from the specialised media monitoring tool enables the observation of situational awareness and the monitoring of changes of chosen aspects within a secure environment.

The remainder of this paper is structured as follows. Section 2 discusses the motivation and purpose of the newly designed specialised media monitoring tool, as well as various design considerations taken to develop the tool. In Section 3, the various components of the tool are identified and described. The benefits and uses of the specialised media monitoring tool are deliberated in Section 4. Section 5 concludes the paper.

2. Background

Open-source information is not what it used to be and the revolution in ICT has fundamentally changed the distribution and accessibility of such information. As of today, open-source information dominates the universe of intelligence analysts and is unlikely to change in the foreseeable future (Gannon, 2014). Using open-source information, analysts can produce intelligence products, which plays an increasingly important role in decision-making processes (OSINT and Intelligence, 2013). To assist with and ensure the timeous delivery of intelligence products, analysts often rely on free or commercial off-the-shelf OSINT tools to improve the efficiency of information retrieval and extraction. Although various OSINT tools exist, the envisioned aim and purpose of such tools may not always effectively address the needs of intelligence analysts. The need for specialised and sophisticated OSINT tools still remains.

The remainder of this section discusses the motivation for the design and development of a new OSINT tool, called the specialised media monitoring tool. Furthermore, the section also highlights the intended purpose of the tool, as well as design considerations, are taken to guide the development of the tool.

2.1 Motivation

The World Health Organization (WHO) on March 11, 2020, declared the novel coronavirus (COVID-19) outbreak a global pandemic (WHO Press Conference on COVID-19, 2020). With the arrival of the COVID-19 pandemic came social changes and nation-wide lockdowns worldwide. Such sudden and radical changes can have a significant impact on human behaviour and societies, as witnessed by the recent Black Lives Matter global movement (Maqbool, 2020). Using the specialised media monitoring tool, digital media can be continuously monitored for public opinion, emerging sensational events, as well as perceiving of national security risks.

Apart from monitoring news from a reactive or sensation standpoint, the monitoring of events and phenomena for trend analysis and other pattern detection, may provide valuable and early detection of changes in the security environment that can be addressed timeously. This may include the detection of fracture points within a specific geographical society on both national and smaller scale, as well as societal concerns that can be identified and monitored for escalation.

Zooming into a practical use case, in Mumbai, law enforcement has set up a social media lab to monitor social media posts for intelligence gathering of the public. Trained policemen took turns in shifts to monitor the social media feeds of the public together with specialised software that flags content that could be potentially troublesome. This reinforces the motivation for the need for an automated media monitoring tool (Trending now: Cops to keep a keen eye on social media, 2013).

Therefore, the specialised media monitoring tool enables continuous observation of situational awareness regarding chosen aspects within a security environment

2.2 Purpose

The immediate challenge for security and business intelligence fraternities is the expanding volume of open-source information. Technological solutions have proven to be part of the answer to the open-source information challenge (Gannon, 2014). There is, however, a need for the intelligence community to invest more in technology that offers powerful processing and analytical capabilities to assess and exploit the vast amounts of open-source information. Currently, various media monitoring and OSINT tools exist to assist analysts with the transition of open-source information into actionable intelligence. Though, as emphasised earlier in the paper, these tools are often limited, or the sheer volume of information collected can influence effective processing. For example, Maltego, developed by Paterva, is a reconnaissance tool for open-source information and allows for the visualisation of the collected information in a graph format. The true power of Maltego lies with the built-in or custom-created transforms, which allows for the analysis of real-world relationships between people, webpages, domains and networks (Maltego, 2020). However, for intelligence analysts, a learning curve might be required to optimally use Maltego, which can in turn limit the adequate use of the tool to obtain intelligence. Another example is Hootsuite, a social media management platform that permits the scheduling, publishing and distribution of content from 35 social media platforms (Hootsuite, 2020). Although Hootsuite consolidates open-source information from multiple social media platforms, the vast amounts of information can overwhelm intelligence analysts.

While the discussions above merely provided a brief critical overview of two existing OSINT tools, it becomes clear that these tools are somewhat rigid in their design. Such inflexibility impacts the efficiency of the OSINT tools to transition open-source information into intelligence. The need for OSINT tools that support adaptability, as well as flexibility concerning open-source information collection and analysis, are clear. The specialised media monitoring tool attempts to fill this void left behind by current OSINT tools. The purpose of the specialised media monitoring tool is to offer versatility using the following strategic features:

- **Monitoring pre-selected digital media platforms:** that offers current, relevant and up to date open-source information that can easily be obtained and analyzed.
- **Perform geographically guide searches for information:** by limiting the amount of open-source information collected to prevent information overload and focus on areas of interest.
- **Direct the extraction of information:** to include elements beyond the obvious, such as text, that can offer insight into situational awareness.
- **Structuring of the collected information:** to enable and allow for quick and easily attainable intelligence.

The various design considerations taken to enable the specialised media monitoring tool to support versatility are captured in the following section.

2.3 Design Considerations

Versatility is a fundamental aspect of the specialised media monitoring tool. Enabling versatility requires the following design considerations to drive the construction of the specialised media monitoring tool (Albin, 2003):

- **Extensibility:** allow for new capabilities to be added to the software without major changes to the underlying architecture.
- **Modularity:** provide well-defined, independent components which leads to better maintainability of the software.
- **Usability:** well-constructed user interface (UI) that is efficient and satisfying to use, as well as easy to learn.
- **Scalability:** permits the software to easily adapt to increasing data, additional functions or more users.
- **Modifiability:** allow the software to grow and change over time.

By guiding the design of the specialised media monitoring using the above-mentioned design consideration, the software tool will be able to grow, adapt and meet the ever-changing requirements of users. Therefore, versatility will be supported, preventing the specialised media monitoring tool from becoming obsolete in the near future. Taking note of these critical design considerations, the following section discusses the various architectural components of the specialised media monitoring tool.

3. Components of the Specialised Media Monitoring Tool

The architectural design of the specialised media monitoring tool consists of the following essential components: (i) data source, (ii) collection methodology, (iii) data storage, and (iv) UI. These components, as well as the flow of interaction, are visualised in Figure 1.

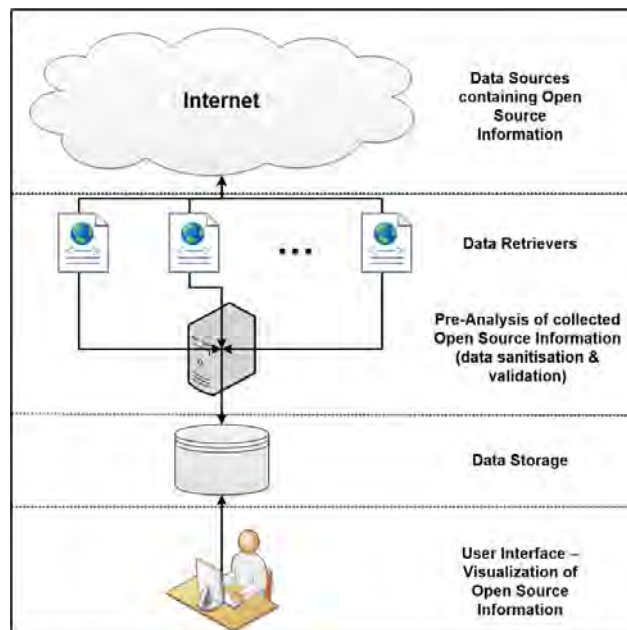


Figure 1: Architectural components of the specialised media monitoring tool

Each component presented in Figure 1 is further discussed and elaborated on in the following sections.

3.1 Selected Data Sources

A significant feature of the specialised media monitoring tool is the monitoring of pre-selected digital media platforms, which must offer current, relevant and up to date open-source information. Based on these requirements, Twitter and online (digital) newspapers were selected as the primary data sources for the specialised media monitoring tool. Twitter is a popular, prominent, and frequently used microblogging and social networking service. The platform offers insight into “what’s happening in the world and what people are talking about right now” (About Twitter, 2020) by sharing messages called tweets. Online or digital newspapers are described as electronic documents that present breaking news promptly. Although various online newspapers do exist, media bias associated with the newspaper must also be taken into consideration. Media bias is defined as “political bias in journalistic reporting, in programming selection, or other in mass communication media” (Definitions for media bias, 2020). Therefore, it is important to ensure that a wide variety of online newspapers and identified open sources are covered in the collection process to compensate for bias and/or validate the information.

Collecting from both Twitter and online newspapers will ensure the specialised media monitoring tool has access to current and relevant open-source information. Although the specialised media monitoring tool currently only collects open-source information from two data sources, the extensibility and modifiability of the architectural design of the tool will permit additional data sources to be included in the future. The following section discusses the collection methodology followed to collect open-source information from the described data sources.

3.2 Collection Methodology

The collection methodology of the specialised media monitoring tool relies on the notion of data retrievers. A data retriever is an automated but dedicated script that enables the collection of open-source information from a specific data source. Data retrievers allow for the continuous collection of open-source information by following a scheduled routine to collect and extract relevant information as per a predefined schema. The specialised media monitoring tool currently consists of four unique data retrievers:

- **Newspaper retriever:** Python-developed script that utilizes the *feedparser* and *newspaper* libraries to collect the latest news articles from various newspapers via Really Simple Syndication (RSS) feeds. The script is scheduled to execute twice daily to ensure complete coverage of the latest news articles.
- **Twitter retriever – location-based:** Python-developed script that uses the *Tweepy* library to access the Twitter Application Programming Interface (API). The script is scheduled to execute hourly, and a predefined list of locations is used to filter the collection of tweets.
- **Twitter retriever – keyword-based:** Python-developed script that uses the *Tweepy* library to access the Twitter API. The script is scheduled to execute hourly, and a predefined list of keywords is used to filter the collection of tweets.
- **Twitter retriever – handle-based:** Python-developed script that uses the *Tweepy* library to access the Twitter API. The script is scheduled to execute hourly, and a predefined list of Twitter handles (individuals of interest) is used to filter the collection of tweets.

The developed data retrievers permit the collection of open-source information from Twitter, as well as local and international newspaper agencies. To enable intelligence extraction from the collected open-source information, the above-mentioned data retrievers must collect more than just the tweet content or the text from a newspaper article. Data elements extracted by the newspaper retriever include the following: newspaper article title, newspaper article text, author of the newspaper article, published date, access date, and Uniform Resource Locator (URL). Data elements extracted by both Twitter retrievers include the following: Twitter handle, tweet, retweet count, “favourited” count, location, hashtags, tweet ID (the unique identifier of the tweet), user ID (the unique identifier of the Twitter user), publish date and access date.

Applying appropriate analysis and data mining techniques to the vast variety of collected data elements will enable the extraction of collated datasets for further analysis towards producing intelligence. Specific and ontology-defined extraction rules can also be set up to create dashboards to enable situational awareness of certain aspects of the security landscape. However, to ensure the accuracy and reliability of the produced intelligence, it is important to ensure the validity of the data used for analysis. To achieve this, data cleansing and validation techniques must be applied to the collected open-source information. Data cleansing is the process of detecting and correcting incomplete, incorrect, inaccurate or irrelevant parts of the collected open-source information (Rahm & Hong, 2000). Deployed data cleansing techniques forming part of the collection methodology include the following:

- Language evaluation and exclusion (currently only analysis of English news articles and tweets supported).
- Removal of incomplete or malformed data elements.
- Consistent formatting and uniformity of date elements to enable periodical searches.

Once the appropriate data cleansing techniques have been applied to the collected open-source information, the cleaned data is transferred to data storage.

3.3 Data Storage Structure

Data storage refers to the collective methods and technologies used to capture and retain digital data. Central to data storage is databases, which allows for the organised collection of data generally stored and accessed electronically (Derclaye, 2005). Although relational databases, which model data as a series of tables consisting of rows and columns (Connolly & Begg, 2003), are still widely used and recommended, a new generation of non-relational databases appeared early in the 21st century. NoSQL (or Not Only SQL) databases attempt to overcome the limitations of a relational database by utilising a distributed, non-rigid and flexible structure to store large amounts of data (Sharma & Meenu, 2012). A well-known example of a NoSQL database is MongoDB.

MongoDB is described as a document-based NoSQL database using a schemaless structure to store large amounts of unstructured data. With MongoDB, the database is a physical container for collections and each collection is a group of MongoDB documents. A document is a set of key-value pairs holding the actual data encoded in a JavaScript Object Notation (JSON) format. The document-oriented model of MongoDB provides key qualities, such as high scalability, heterogeneity and improved performance (Krishnan, Sudheep Elayidom, & Santhanakrishnan, 2016).

The obtained and pre-processed (cleaned) open-source information to be retained in data storage is not organised according to a pre-defined model. The unstructured nature of the open-source information is due to the different data sources used for the collection of the open-source information, as well as the data cleansing techniques deployed to remove unusable data elements. Therefore, the specialised media monitoring tool requires a database structure capable of handling data with a dynamic schema. The key qualities of MongoDB, as well as the ability to support advanced queries and perform complicated operations, confirm the appropriateness of MongoDB as the database for the specialised media monitoring tool.

The visualisation of the open-source information captured and stored in the database is further discussed in the following section.

3.4 User Interface Design

The intended purpose and use of the specialised media monitoring tool are to assist analysts to effectively observe situational awareness. Therefore, the UI of the specialised media monitoring tool is wisely crafted to enable usability and permit users to extract the necessary intelligence from the collected open-source information. Figure 2 presents a visual representation of the UI design for the specialised media monitoring tool.

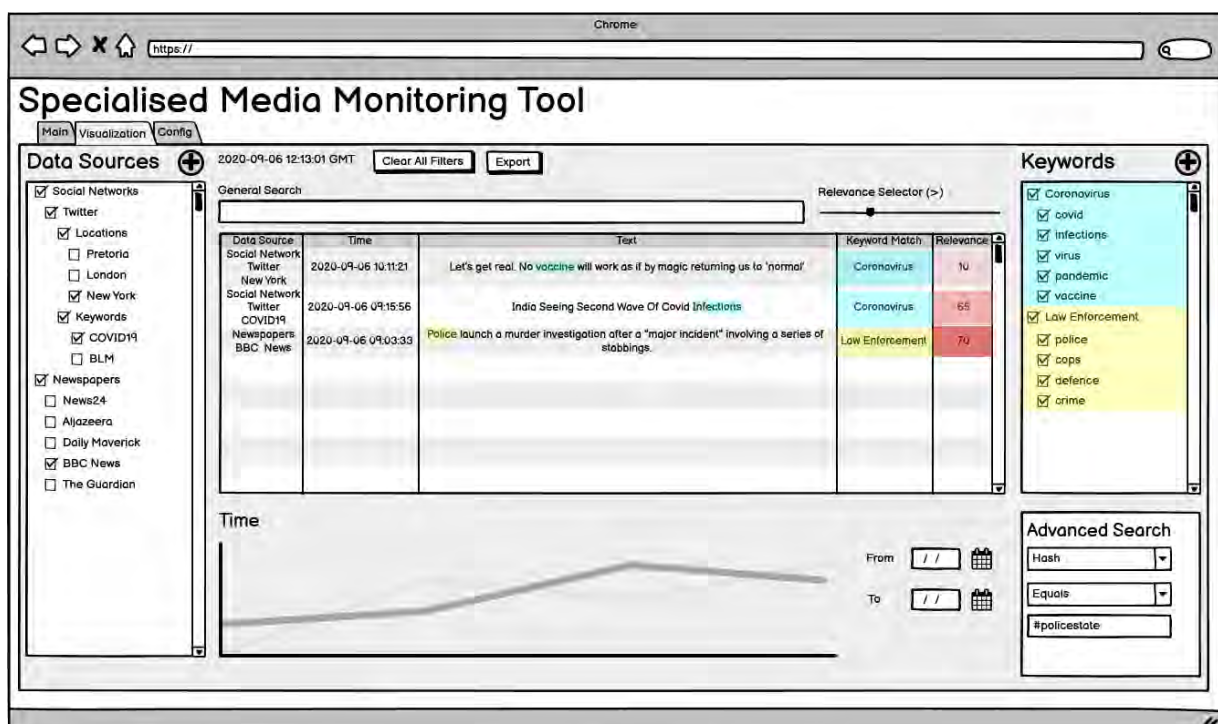


Figure 2: UI of the specialised media monitoring tool

The UI design of the specialised media monitoring tool includes the following key UI elements to allow the user to select and control the display of open-source information via the UI:

- **Data source selection:** choosing from the available data sources, either primary (Twitter or online newspapers) or pre-analysed subsets extracted from the primary data sources. The selection of the data sources will limit the amount of open-source information displayed and processed during searches.
- **General searches:** conduct a non-specialised search to find results using the selected data sources.
- **Pre-defined keyword searches:** conduct specialised searches by selecting pre-defined keywords from set categories (e.g., Coronavirus or Law Enforcement) to find results using the selected data sources.
- **Advanced searches:** conduct specialised searches by constructing unique queries to search for specific results from the selected data sources.
- **Date-based search:** conduct specialised searches using date filters to limit the amount of open-source information displayed and processed during searches.
- **Relevance selection:** limit the amount of open-source information displayed based on the quality of the information.
- **Clear all filters:** reset to default selections and clear all results.

- **Exporting results:** produced by searches and selected filters to allow for offline analysis.
- **Visualisation of results:** produced by searches and selected filters to simplify observation of situational awareness (illustrated in Figure 3). Visualisation techniques include
 - Word clouds – a graphical representation of all words contained within the selected data sources, in which the size of each word indicates its frequency.
 - Sentiment analysis – categorising the emotion, either positive or negative, of the selected data sources.

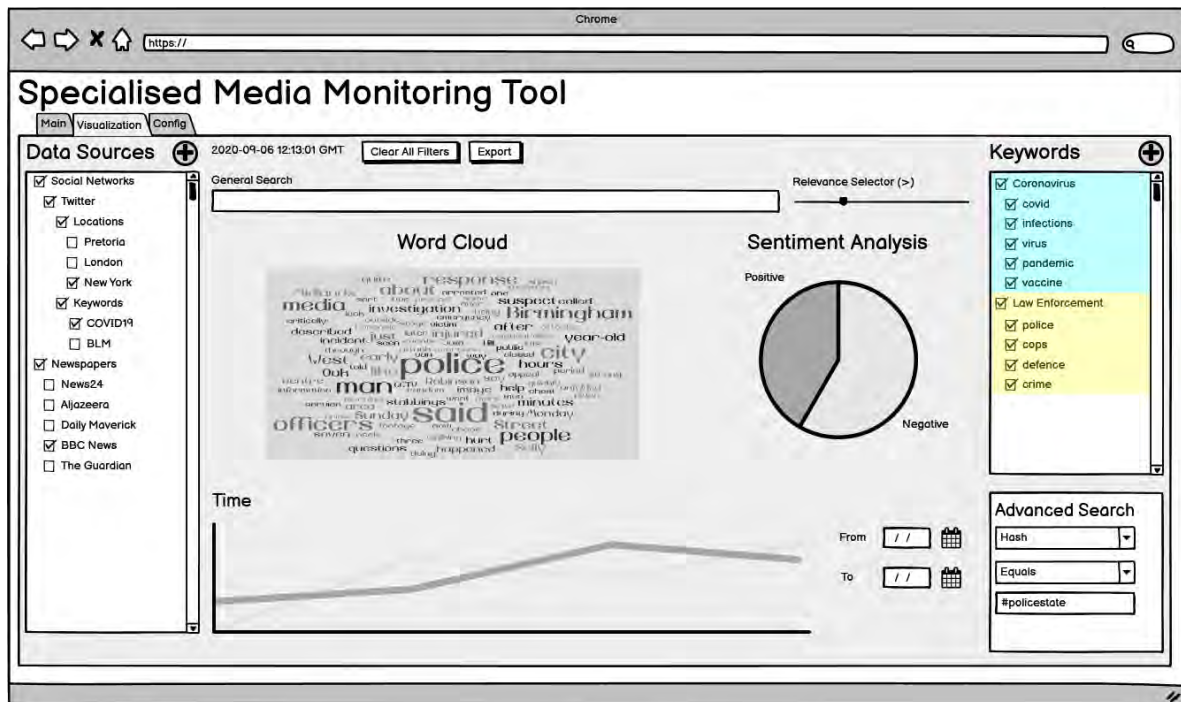


Figure 3: Visualisation of techniques of specialised media monitoring tool

The purpose of the above UI elements is to enable the usability of the specialised media monitoring tool, allowing users to monitor situational awareness in near real-time.

4. Benefits of Specialised Media Monitoring Tool

The specialised media monitoring tool is developed specifically to assist analysts to more effectively collect and process open-source information, which enables the observation of situational awareness. The primary benefit of the specialised media monitoring tool is that it assists the military or business intelligence functionary in the collection and pre-analysis of data towards producing good quality intelligence products. The Intelligence Cycle describes a set of functions of the processes used to create intelligence to enable improved decision-making. The cycle starts with a problem and consists of several processes including planning & direction, collection, processing, analysis & production, dissemination & utilisation/integration (Johnson, 1986). The specialised media monitoring tool, therefore, supports the first three processes of the Intelligence cycle.

- It enables data collection from online platforms.
- It enables directed data collection and thus data storage of only relevant data.
- Enables data mining using specialised searches.
- Enables pre-analysis by tagging keywords.
- User-friendly and client customised user interface.

The obtained intelligence products from digital platforms, such as social media and online newspapers, can enable security agencies (e.g., law enforcement or the military) to closely observe the situational awareness of a security environment. Continuous situational awareness allows for the detection of security incidents or triggers that may result in dangerous events.

A key example of the value the specialised media monitoring tool can add to the security agencies is within the capacity of early warning or profiling. Analysts could detect troublesome messages on social media platforms

and act upon them before they cause damage. An example of using social media to monitor events for situational awareness is the 2019 Christchurch shooting where two consecutive mass shootings occurred at mosques in a terrorist attack in Christchurch, New Zealand, during prayers. Parts of the mass shooting were live streamed on social media. Furthermore, a hate-filled document was also published some 10 to 20 minutes before the attack. Had law enforcement agencies been monitoring social media, such an event could have perhaps been avoided due to early warning.

Another benefit of the specialised media monitoring tool is the ability to simultaneously monitor a range of threat scenarios, such as civil unrest, rival factions (e.g., gangster groups or crime syndicates), terrorist activities, and political developments. For example, protests are usually coordinated and shared on social media platforms. Law enforcement agencies that monitor these could warn people about safety in various locations. Furthermore, analysing these social media posts could provide insights into causes for such events and key role players who require further investigation.

Finally, the features incorporated into the specialised media monitoring tool enables embassies to utilise the tool to observe country risks, political developments within a nation, as well as civil society's reaction to government initiatives.

5. Conclusion

This paper aimed to introduce the new specialised social media monitoring tool, which was designed to assist analysts to observe situational awareness. The specialised media monitoring tool guides the collection of open-source information to enable the creation of intelligence products by incorporating four distinct features: (i) monitoring pre-selected digital media platforms, (ii) geographically guide searches for information, (iii) direct the extraction of information, and (iv) structuring of the collected information. By making use of the specialised media monitoring tool, analysts can monitor open-source information, which is collected from online newspapers and social media feeds, to attain intelligence. The obtained intelligence, thus, enables the observation of situational awareness of chosen aspects within a security environment, providing analysts with insights into the causes and details of security events. Although valuable information for situational awareness is contained within the open-source data sources, it is still a complex process to extract relevant information and to detect useful patterns in the data to identify threats and to predict possible attacks. Automation is required for intelligent decision-making. For future work, the authors will investigate the use of semantic technologies to enable the software system to further process the data and annotate it with meaning.

Acknowledgements

The authors would like to thank Roelof Temmingh for sharing his technical insight, guidance and support during the design and development of the specialised social media monitoring tool.

References

- About Twitter*. (2020). Retrieved from Twitter: <https://about.twitter.com/>
- Albin, S. T. (2003). *The Art of Software Architecture*. Indianapolis: Wiley Publishing Inc.
- Best, C. (2011). Challenges in Open Source Intelligence. European Intelligence and Security Informatics Conference.
- Connolly, T., & Begg, C. (2003). *Database Systems*. Pearson.
- Definitions for media bias*. (2020). Retrieved from Definitions: <https://www.definitions.net/definition/media+bias>
- Derclaye, E. (2005). What is a Database. *The Journal of World Intellectual Property*, 5(6).
- Forsyth, A. (1999). The Information Age: Economy, Society and Culture. *Journal of Planning Education and Research*, 19(2).
- Gannon, J. (2014). The Strategic Use of Open-Source Information. *Intelligence Community Perspective*, 67-71.
- Gheisari, M., Irizarry, J., & Horn, D. B. (2010). Situation awareness approach to construction safety management improvement. *Proceedings of 26th Annual ARCOM Conference* (pp. 311-318). Leeds, UK: Association of Researchers in Construction Management.
- Haris, A. R. (2016). Information Issues in Digital Era. Faculty of Information Management, Universiti Teknologi.
- Hootsuite*. (2020). Retrieved from Hootsuite: <https://hootsuite.com/>
- Johnson, L. K. (1986). Making the intelligence "Cycle" work. *International Journal of Intelligence and Counter Intelligence*, 1(4), 1-23.
- Krishnan, H., Sudheep Elayidom, M., & Santhanakrishnan, T. (2016). MongoDB - a comparison with NoSQL databases. *International Journal of Scientific & Engineering Research*, 7(5), 1035-1037.
- Maltego*. (2020). Retrieved from Maltego: <https://www.maltego.com/>
- Maqbool, A. (2020, July 10). *Black Lives Matter: From social media post to global movement*. Retrieved from BBC News: <https://www.bbc.com/news/world-us-canada-53273381>

- Noble, D. F. (2004). Assessing the Reliability of open Source Information. Stockholm, Sweden: Proceedings of the 7th International Conference on Information Fusion.
- OSINT and Intelligence. (2013). *Reuser's Information services*, 1-8.
- Rahm, E., & Hong, H. D. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*, 23, 3-13.
- Sharma, V., & Meenu, D. (2012). SQL and NoSQL Databases. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(8), 20-27.
- Trending now: Cops to keep a keen eye on social media*. (2013, March 17). Retrieved from The Indian Express: <http://archive.indianexpress.com/news/trending-now-cops-to-keep-a-keen-eye-on-social-media/1089333>
- WHO Press Conference on COVID-19*. (2020, March 11). Retrieved from WHO: https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-and-final-11mar2020.pdf?sfvrsn=cb432bb3_2
- Williams, H. J., & Blum, I. (2018). *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise*. Santa Monica: RAND Corporation.

A New Dawn for Space Security

Jordan J. Plotnek and Jill Slay

University of South Australia, Adelaide, Australia

Jordan.Plotnek@mymail.unisa.edu.au

Jill.Slay@unisa.edu.au

Abstract: Space infrastructure provides vital services for a number of critical industries, including; defence, transportation, energy, utilities, emergency services, banking, environment, academia, and others. These services range from global communications to remote sensing and geolocation, with many new applications undoubtedly on the horizon, including plans for further exploration and even human settlement. It is therefore essential that space technologies are protected from unwanted interferences – a task that is becoming more challenging by the day. Adding to the already complex space security environment, we are experiencing the beginnings of a second space race that is seeing the rapid deployment of space systems containing a vast array of new technologies, such as the Internet of Things (IoT) and advanced onboard processing. This is subsequently introducing new vulnerabilities to an already aged and vulnerable satellite ecosystem, hence increasing the risk of potentially catastrophic security events. Although well-articulated in political, legal, and international relations literature, the engineering, science, and technology aspects of space security are currently under-studied and disjointed, leading to fragmented research and inconsistent terminology. This paper examined space security from an engineering perspective by conceptually tying existing space and security literature together to detail the space threat landscape and identify research gaps and opportunities. Additionally, this paper identifies the need for wider recognition of space systems security as a specialist inter-disciplinary domain in order to break down disciplinary silos, enhance collaboration, and unify definitions, taxonomies, and research objectives.

Keywords: critical infrastructure, cyber threat, resilience, satellite, space security, space weapon

1. Introduction

In 1957 the Soviet Union launched Sputnik-1, the first manmade object to enter earth's orbit, triggering a two-decade long space race and forever changing the course of human history. Today we are on the cusp of a second space race, this time with almost 3000 artificial satellites already in operation (UCS, 2020), and countless more space debris littered in perpetual orbit. As the battle for space superiority ramps up for a second time, we are forced to acknowledge the wildly different technological landscape compared to that of our scientific colleagues back in 1957.

Industrial systems are increasingly both particularly vulnerable and specifically targeted by adversarial groups and state actors (Kaspersky Lab, 2019), and space systems are not immune. In fact, as detailed in the following sections, space systems face an even greater range of threats and have even further reaching consequences than those combatted by terrestrial critical infrastructures (Bradbury et al., 2020), and considering the level of military dependence on space infrastructure and the volatile state of global affairs today (Donnelly, 2021), there has never been a more urgent need for resilient space systems as there is now.

This paper aims to lay the academic foundations in response to the existing void in literature on security and resilience in a space systems context. The paper commences with a review of published literature in the field at Section 2, moving onto setting the scene by introducing space infrastructure, space threats, past events, and future trajectories in Section 3. With context in hand, the need for space systems security as a recognised interdisciplinary domain is then presented in Section 4, followed by a discussion of potential space systems security research opportunities in Section 5 and finally concluding in Section 6.

2. Foundational Literature

Of the various disciplines contributing to space security knowledge, the social sciences are the most mature, with several decades of published history. Traditionally, space security has been viewed primarily as a military domain due to Cold War motivations behind the first space race (Sheehan, 2015). More recently, however, this view has expanded to include three dimensions of space security (Mayence, 2010):

1. security in space (i.e., protecting space systems);
2. space for security (i.e., military space operations); and
3. security from space (i.e., protecting Earth from space-based threats).

This paper focuses exclusively on the first dimension of space security, herein referred to as 'Space Systems Security'. Drawing from older literature we can find several space security definitions that can be reapplied directly to space systems security. Moltz's definition serves as a good baseline, defining it as "the ability to place and operate assets outside the Earth's atmosphere without external interference, damage, or destruction" (Moltz, 2011).

A whitepaper by the United States Office of the Assistant Secretary of Defense for Homeland Defense & Global Security entitled 'Space Domain Mission Assurance: A Resilience Taxonomy' (USDoD, 2015) gets particular attention amongst space resilience advocates, however it does little to set the scene before proposing a resilience taxonomy that is detached from tangential resilience and security literature.

A second key text in this domain is the book entitled 'Critical Space Infrastructures: Risk, Resilience and Complexity' (Georgescu et al., 2019), which successfully introduces space system fundamentals and examines space systems as critical infrastructure but is decidedly lacking in its discussion of cyber security issues. Though, the introduced taxonomy is helpful as it splits Critical Space Infrastructure (CSI) into five distinct categories, as well as discussing interdependencies with other systems like those for water and energy.

Harrison et al. conducted a 'Space Threat Assessment' in 2020 (Harrison et al., 2020) that focuses on the threat of counter-space weapons, breaking them down into four broad but useful categories: kinetic physical, non-kinetic physical, electronic, and cyber. The remainder of the report is less repurposable and goes on to analyse different nation state capabilities and their threat to the United States at the point in time of the assessment.

A paper entitled 'Cybersecurity Threats to Satellite Communications' (Housen-Couriel, 2016) establishes a typology of state actor responses. However, it does not adequately address space security from a technical perspective. In contrast to Harrison et al., Housen-Couriel identifies only three kinds of satellite 'disruptions': kinetic (direct impact of one satellite with another), virtual (interference with communications), and hybrid (electromagnetic pulse, or EMP, weapons). It then plots these three disruption categories against five stages of satellite operations:

1. pre-launch;
2. at launch;
3. telemetry, tracking, and command (TT&C);
4. transmissions; and
5. end-of-life.

Livingstone and Lewis (2016) take a high-level approach to space cybersecurity, discussing topics such as cyber threats and risks to satellite infrastructure, as well as challenges and trends in the industry. However, the paper is directed toward a general audience so is not guided by existing taxonomies, and hence does not serve the purpose of a foundational academic text. It also is limited to cyber threats alone, which forms only one threat type that a space security practitioner must be aware of.

Finally, a comprehensive paper by Pavur and Martinovic (2020) details the cybersecurity threats to satellites and examines over 100 significant satellite hacking incidents over the past 60 years. The paper identifies four sub-domains that satellite cybersecurity applies to: satellite radio-link security, space hardware security, ground station security, and operational/mission security. They comment on the cross-disciplinary nature of space security but, perhaps due to their narrow focus on cybersecurity, stop short of treating space security as a separate domain in its own right.

A few other papers touch on the subject (Hannan, 2018; Ikitemur et al., 2020; Kallberg, 2012; Kang et al., 2018; Santamarta, 2014) but are specific to niche technologies or formal methods and hence do not adequately lay the foundations for future research on space systems security as a holistic domain. It is important to note that this literature review was only conducted across open-source English resources, not only skewing the threat context to a Western bias, but also excluding any additional or conflicting research that may exist within classified archives.

3. The Space Context

Space is the next frontier for human civilisation. Humans have long relied on space infrastructure for the advancement of technologies here on earth, with such dependencies becoming more and more critical. We are now even on the path toward developments such as extra-terrestrial colonisation, space mining, and other feats that were unimaginable a mere three generations ago.

3.1 Space Infrastructure

CSI can be broken down into five key categories (Georgescu et al., 2019):

- Remote Sensing;
- Communications;
- Meteorological;
- Global Navigation Satellite Systems (GNSS); and
- Administrative and Legislative Frameworks.

The technologies covered by the above are predominantly artificial satellites, but may also include space stations, rovers and vehicles, rockets, space probes, ground stations, and terrestrial communications links. Naturally each of these systems have various unique processes, technologies, and vulnerabilities.

Remote sensing involves the passive or active collection of data about a subject of study without making physical contact. Space infrastructures that fall under this category include systems that conduct surveillance, scientific monitoring, or information gathering for things like terrain mapping and military reconnaissance. These kinds of systems are particularly vulnerable to laser attacks as they allow for electromagnetic penetration to achieve their primary function (Georgescu et al., 2019).

Communications Satellites (ComSat) provide global telecommunications coverage and are useful for aviation and long-distance connections where earth's curvature inhibits the line of sight; communications which, if interrupted, could result in significant loss of life here on earth. A study done by Steinberger at the US Joint Electronic Warfare Center found that the most vulnerable component of satellite communications infrastructure is the antenna, which exposes the satellite to attacks such as jamming or spoofing (Steinberger, 2008). The earth segment was also found to be particularly vulnerable to jamming, but also to threats stemming from internet connectivity.

Meteorological space infrastructures are generally used to monitor earth's climate and weather and are critical for tasks like extreme weather prediction and monitoring. These satellites are generally quite minimal in build, as their primary purpose is simply to transmit photos and meteorological data to earth. There is yet to be any published research specific to meteorological satellite vulnerabilities, however due to their simple anatomy it can be inferred that they likely share general commonalities with other satellite systems.

GNSS includes navigation, positioning, and timing applications, and is perhaps most recognisable in satellite technologies such as the Global Positioning System (GPS). GNSS are heavily relied on by terrestrial applications such as the electric grid and guided weapons systems, whereby a satellite failure could cause far-reaching and catastrophic consequences, including loss of life. Due to the relatively long history of such systems, satellites delivering GNSS capabilities have been privy to greater levels of security research compared to other space-based systems. Across the literature, jamming and spoofing surface as the primary vulnerabilities of GNSS (Ioannides et al., 2016; de Abreu Faria et al., 2016; Amin et al., 2016).

Administrative and Legislative Frameworks are undoubtedly a quintessential component of CSI and are also notably immature at this point in time (Planck, 2009). A growing number of countries around the world are recognising space systems as Critical National Infrastructure (CNI) and hence the administrative and legal frameworks to support them are gaining global attention and prioritisation. One notable example is The Woomera Manual project, which is an international research collaboration to articulate existing international laws applicable to military space operations (Stephens, 2021). Space as a legal domain is notoriously complex due to its international significance and lack of any divisible territory (del Monte, 2013). For space security purposes this category functions as more of a supporting component because, although frameworks are not targetable by a threat actor, they can aid in pre-emptive security efforts, data collection and retention standards, post-compromise forensics, and attribution, prosecution and retaliation.

3.2 Threat Landscape

Space systems operate in one of the most naturally hostile environments known to man, constantly facing threats such as electromagnetic radiation and space debris. In addition, systems deployed in space also face a variety of unique challenges that don't commonly apply to terrestrial infrastructure, such as lack of redundancy or maintenance options (Georgescu et al., 2019). Although non-malicious threats should definitely be considered when risk assessing space technologies, this is outside the scope of this paper.

When discussing targeted threats, it is helpful to break them down into three components (see Fig. 1): the actor, the vector, and the attack. The threat actor is the person or organisation behind the attack and can be assessed by considering their capability to conduct an attack versus their intent behind the attack. The threat vector refers to the vulnerable point of entry used by the threat actor to successfully carry out the attack; for example, if a ground system is air gapped (i.e., not connected to any network) then the threat vector may be a flash drive. Finally, the attack itself is the exploit used by the threat actor to achieve their objectives and cause the desired impact, for example malware or spoofing.

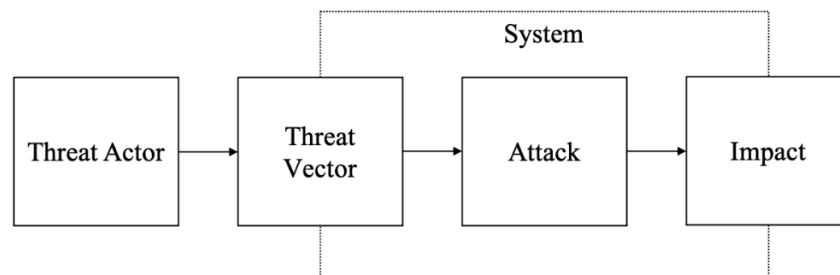


Figure 1: Anatomy of a targeted threat

Although a formal threat actor taxonomy is yet to emerge in the literature, a threat actor is generally categorised as one of the following: nation-state, terrorist, criminal group, or individual (e.g., insider threats, hacktivists, or bored teenagers) (Livingstone and Lewis, 2016). Sometimes hacktivism occurs on a larger scale (e.g., Anonymous) and can be treated as a separate category, conforming to neither terrorist nor criminal motivations. Bradbury et al. broke these high-level categories down further and provided space-specific examples, whereby each space-cyber actor has their own intent (e.g., goals and motivations) and capability (e.g., capabilities, environment, and resources) that drives their decision-making process when carrying out a targeted attack (Bradbury et al., 2020). Pavur and Martinovic produced a similar yet simpler version of this threat actor table, expanding beyond just cybersecurity considerations in Figure 2.

Attacker Type	Example Motivations	Technical Capabilities
National Military	- Space Control - Anti-Satellite Weapons	High
State Intelligence	- Counter-Intelligence - Technology Theft - Eavesdropping	High
Industry Insiders	- Sabotage - Technology Theft	High
Parts Suppliers	- Sabotage - Espionage	High
Organized Crime	- Eavesdropping - Ransom - Technology Theft	Moderate
Commercial Competitors	- Sabotage - Technology Theft	Moderate
Terrorists	- Societal Harm - Notoriety - Message Broadcasting	Low
Individuals	- Notoriety - Personal Challenge	Low
Political Activists	Message Broadcasting	Low

Figure 2: Summary of Satellite Threat Actors by Pavur and Martinovic, 2020

Threat vectors need to be assessed on a case-by-case basis, however there are four common attack surfaces for deployed space systems (Wheeler et al., 2018):

- inputs (e.g., sensors and RF antennae);
- outputs (e.g., telemetry transmitters);
- internal communications (e.g., Spacewire buses); and
- computing (e.g., the internal system that integrates each component).

Each of these components can be accessed via a myriad of different threat vectors, such as through ground segments, supply chains, unsecured communications links, and countless other avenues. Bradbury et al. propose a handy reference architecture for assessing space system threat vectors and attack surfaces in their 2020 IEEE Aerospace Conference paper (Bradbury et al., 2020).

Targeted attacks to space infrastructure can be broken down into: kinetic physical, non-kinetic physical, electronic, and cyber; as per the threat assessment published by Harrison et al. (Harrison et al., 2020). In this context both kinetic and non-kinetic physical threats aim to impact the physical components of a space system. The difference between the two is inherent in the title, with kinetic referring to tangible threats such as anti-satellite (ASAT) missiles and non-kinetic referring to intangible threats such as lasers and EMP weapons. It is worth noting here that kinetic weapons are particularly risky as any ensuing space debris can cause cascading failures, where one collision leads to the next and suddenly a large number of satellites are transmorphed into space junk; including unintended targets (Wright et al., 2005).

Electronic threats do not aim to have a permanent physical impact, and so are not to be confused with non-kinetic physical threats that do. An electronic threat generally involves interfering with RF signalling, for example signal jamming or spoofing, to interfere with the availability or integrity of communications, with the consequences to the space infrastructure itself usually being temporary.

Finally, cyber threats seek to interfere with the confidentiality, integrity, or availability of space infrastructures through the manipulation of data and code. Cyber threats are the most flexible of the categories, with a wide range of malicious options and outcomes available to the adversary (Plotnek and Slay, 2021a). Cyber attacks are rapidly growing in occurrence and severity due to their accessibility, affordability, and the increased level of control an actor has over the impact compared to alternative forms of attack. As such, cyber attacks deserve special attention when researching threats to space systems. The Cyber Kill Chain is a conceptual model invented by Lockheed Martin to understand the various stages of a cyber attack (i.e., reconnaissance, weaponization, delivery, exploitation, installation, command & control, and actions on objectives), and helps to analyse attacks and attack vectors for prevention and incident response. In their 2021 conference paper Van der Watt and Slay adapted the Cyber Kill Chain model to Low Earth Orbit (LEO) satellites; a helpful reference when discussing cyber threats to space infrastructure (van der Watt and Slay, 2021).

3.3 Past Events

Unfortunately, secrecy clouds public access to information about incidents to critical infrastructure, leaving the idea of space systems security to the imagination of Hollywood. The lack of academic research in the field only exacerbates the fallacy that space security is a hypothetical need. Pavur and Martinovic (2020) conducted a study of historical satellite hacking incidents and identified 116 significant events since Sputnik, with the first occurring in 1986.

The first few years of space attacks saw a heavy focus on piracy and spoofing, with satellite imagery data being eavesdropped to avoid subscription fees and television streams being hijacked to broadcast unsolicited messages. A noteworthy example is the 1987 hack conducted by an employee of the American Christian Broadcasting Network who transmitted unauthorised biblical messages over the Playboy Channel's planned broadcast. The 1990s saw a move towards signal jamming, with commercially available satellite jammers being produced and state actors such as the US, Iran, Indonesia, and Russia carrying out various jamming operations.

The use of commercial and state-sponsored jamming, signal hijacking, laser attacks, malware, eavesdropping, and other increasingly sophisticated attacks have been becoming more and more common (Pavur and Martinovic, 2020). In 2007 China even compromised two NASA satellites via the ground station, taking complete control over their flight signalling (Bardin, 2013). That same year China also demonstrated a kinetic ASAT weapon

against one of their own satellites, producing hundreds of pieces of dangerous space debris along with it, and playing a role in the onset of the second space race that we are witnessing today (Zissis, 2010).

There are many more examples of malicious threats to space infrastructure that solidify the need for space security. Appendix A of Pavur and Martinovic's paper contains a comprehensive list of past cyber attacks on satellite infrastructure (Pavur and Martinovic, 2020).

3.4 Looking Towards the Future

Where the first space race cemented space systems as critical infrastructure, the second space race is shifting the focus from government to commercial interests. According to Livingstone and Lewis (2016), the next decade or so will see system-on-a-chip avionics, self-optimizing autonomous systems, complex on-board satellite processing, autonomous satellite-to-satellite (S2S) communications, plus a number of complex software additions. Each technological advancement introduces new vulnerabilities that could be exploited, producing unseen effects. For example, consider a futuristic piece of worm-like malware that corrupts a satellite connected via an autonomous S2S system – the entire fleet could be compromised and potentially rendered unserviceable after a single infection.

Alongside this resurgence in the rapid development of space systems, all kinds of new threats are emerging. Talk of cyber warfare, cyber terrorism, and cyber crime are increasing and so are the capabilities of motivated threat actors (Plotnek and Slay, 2021b). Both cyber and electronic weapons are becoming more effective and accessible by the day, with at least 120 different countries already invested in cyber warfare capabilities (McAfee, 2005).

Mass-scale environmental and political events may also impact humankind's reliance on CSI, which could cause unforeseeable impacts. For example, hazardous asteroids heading for earth (O'Neill and Handal, 2021) or the growing threat of climate change, both of which are tracked and assessed using space infrastructure – a reliance that may evolve and become more critical as time goes on. Another example might be a third eruption of world warfare. Military equipment has become increasingly reliant on satellite technology and such a situation may over-burden aging infrastructure and cause denials of service in critical moments. On a similar tangent, the United States has officially approved the establishment of a Space Force to directly counter these threats (Farley, 2020) and many other countries are likely to follow suit, events that will undoubtedly impact the space security domain in the coming decades.

4. The Case for Space Systems Security

With an understanding of the criticality of space infrastructure, its deepening vulnerability issues, and the unpredictable threat environment within which it is situated, it is easy to see the importance of space security. Unfortunately, up until now there has been little recognition or structure afforded to the complex domain of space systems security. In fact, there even exists some hostility towards security research within the space industry (CSRIC, 2015), a culture that could impede the advancement of space development altogether.

The second space race has sparked a period of rapid development and deployment, which presents significant complications without a unified understanding of the domain's research problems for efficient prioritisation and collaboration. The current lack of direction and common purpose has led to a double-up in the limited research available, with each contributing discipline evidently taking a siloed approach to space security terminology and taxonomy. Additionally, unlike a lot of other critical infrastructures space has direct military applications, meaning that efficient research and development is crucial for national security objectives such as effective threat deterrence and space superiority.

The kind of efficiency needed to compete in the volatile arena of this new space race is only made possible through wider recognition of space systems security as a specialist inter-disciplinary domain, where each contributing field has a valid voice for enhanced collaboration.

5. Research Opportunities

Pulling together the different research threads identified in Section 2 paints a picture of the current state of space security as a domain and highlights areas needing further development, as shown in Figure 3. As can be seen, we have a preliminary understanding of attack surfaces, threats, and actors, as well as past events and

future predictions, which are not shown in the diagram. Findings in these areas should be verified and formally defined and taxonomised to ensure cross-disciplinary consistency and to aid collaboration.

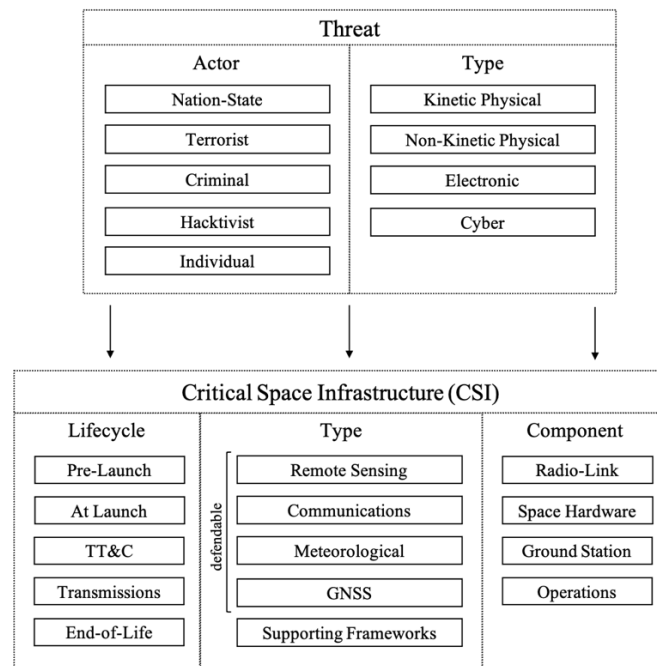


Figure 3: Threats to CSI broken down into taxonomical sub-categories as per available literature

Expanding on Harrison et al. we can class malicious (i.e., non-environmental) space threats under four categories: kinetic physical, non-kinetic physical, electronic, and cyber (Harrison et al., 2020). These four categories are more descriptive for general security use compared to the three law-driven categories (kinetic, virtual, and hybrid) proposed by Housen-Couriel in their earlier paper (Housen-Couriel, 2016), however more research may be required to determine a universally robust space threat taxonomy. In that same paper, Housen-Couriel identifies five stages of satellite operations, which should be confirmed by satellite engineering academia and analysed from a mission security perspective.

With a clear idea of the current state of literature it becomes easier to see where research should be focused in future. Firstly, space systems security should be formally defined and taxonomised according to the in-scope domains or high-level sub-topics. A research roadmap should then be developed that identifies the fundamental objectives of space systems security as a formal domain, the key stakeholders to the problem of space systems security (i.e., contributing specialisations), and how each stakeholder interacts across each taxonomical sub-domain and objective.

6. Conclusions

Far from a hypothetical consideration, vulnerable space infrastructures have already experienced a plethora of significant security events and this will only increase in frequency and impact looking towards the high-tech and interconnected future of space flight. Earth-based societies are already dependent on space infrastructure and face dire consequences without the proper consideration of space systems security.

This paper has established a contextualised foundation for space systems security drawn from existing cross-disciplinary literature on the subject. Additionally, it was found that wider recognition of space systems security as a specialist inter-disciplinary domain is needed to break down disciplinary silos, enhance collaboration, and unify definitions, taxonomies, and research objectives. Finally, the paper painted a picture of the current state of space systems security research and drew from this to propose a direction for future research.

References

- Amin, M.G., Closas, P., Broumandan, A., Volakis, J.L., 2016. Vulnerabilities, threats, and authentication in satellite-based navigation systems [scanning the issue]. *Proceedings of the IEEE* 104, 1169–1173. <https://doi.org/10.1109/JPROC.2016.2550638>

- Bardin, J., 2013. Satellite Cyber Attack Search and Destroy, in: *Computer and Information Security Handbook*. Elsevier Science & Technology, pp. 1093–1102.
- Bradbury, M., Maple, C., Hu, Y., Ugur, I.A., Cannizzaro, S., 2020. Identifying Attack Surfaces in the Evolving Space Industry Using Reference Architectures, in: 2020 IEEE Aerospace Conference. Presented at the 2020 IEEE Aerospace Conference, IEEE, USA, pp. 1–20. <https://doi.org/10.1109/AERO47225.2020.9172785>
- CSRIC, 2015. Cybersecurity Risk Management and Best Practices Working Group 4: Final Report (Working Group No. 4). The Communications Security, Reliability and Interoperability Council, U.S.A.
- de Abreu Faria, L., de Melo Silvestre, C.A., Correia, M.A.F., 2016. GPS-Dependent Systems: Vulnerabilities to Electromagnetic Attacks. *Journal of Aerospace Technology and Management* 8, 423–430. <https://doi.org/10.5028/jatm.v8i4.632>
- del Monte, L., 2013. Towards a cybersecurity policy for a sustainable, secure and safe space environment, in: *Proceedings of the 64th International Astronautical Congress (IAC)*.
- Donnelly, D., 2021. War fears as China vows punishment on Australia for Taiwan support - Long-range strikes. *Express*.
- Farley, R., 2020. Space Force: Ahead of Its Time, or Dreadfully Premature? CATO Institute Policy Analysis 904.
- Georgescu, A., 2020. Critical Space Infrastructures, in: Schrogl, K.-U., Hays, P.L., Robinson, J., Moura, D., Giannopapa, C. (Eds.), *Handbook of Space Security*. Springer, New York, United States of America, pp. 227–244.
- Georgescu, A., Gheorghe, A.V., Piso, M., Katina, P.F., 2019. Critical Space Infrastructures: Risk, Resilience and Complexity, *Topics in Safety, Risk, Reliability and Quality*. Springer, Switzerland.
- Hannan, N., 2018. An Assessment of Supply-Chain Cyber Resilience for the International Space Station. *The RUSI Journal* 163, 28–32. <https://doi.org/10.1080/03071847.2018.1469249>
- Harrison, T., Johnson, K., Roberts, T.G., Way, T., Young, M., 2020. Space Threat Assessment 2020. Center for Strategic & International Studies, Washington, United States.
- Housen-Couriel, D., 2016. Cybersecurity threats to satellite communications: Towards a typology of state actor responses. *Acta Astronautica* 128, 409–415. <https://doi.org/10.1016/j.actaastro.2016.07.041>
- Ikitemur, G., Karabacak, B., Igonor, A., 2020. A Mixed Public-Private Partnership Approach for Cyber Resilience of Space Technologies, in: *Space Infrastructures: From Risk to Resilience Governance*, NATO Science for Peace and Security Series - D: Information and Communication Security. IOS Press, pp. 120–130.
- Ioannides, R.T., Pany, T., Gibbons, G., 2016. Known Vulnerabilities of Global Navigation Satellite Systems, Status, and Potential Mitigation Techniques. *Proceedings of the IEEE* 104, 1174–1194. <https://doi.org/10.1109/JPROC.2016.2535898>
- Kallberg, J., 2012. Designer Satellite Collisions from Covert Cyber War. *Strategic Studies Quarterly* 6, 124–136.
- Kang, M., Hopkinson, K., Betances, A., Reith, M., 2018. Mitigation of Cyber Warfare in Space Through Reed Solomon Codes. Presented at the 13th International Conference on Cyber Warfare and Security, acpi, Washington DC, United States.
- Kaspersky Lab, I.C., 2019. Threat Landscape for Industrial Automation Systems. Kaspersky Lab, ICS CERT.
- Livingstone, D., Lewis, P., 2016. Space, the Final Frontier for Cybersecurity? Chatham House.
- Mayence, J.-F., 2010. Space security: transatlantic approach to space governance, in: Robinson, J., Schaefer, M., Schrogl, K.-U., von der Dunk, F. (Eds.), *Prospects for Transparency and Confidence-Building Measures in Space*. ESPI, Vienna, Austria, p. 35.
- McAfee, 2005. McAfee Virtual Criminology Report: North American Study into Organized Crime and the Internet. McAfee.
- Moltz, J.C., 2011. *The Politics of Space Security: Strategic Restraint and the Pursuit of National Interests*, Second. ed. Stanford University Press, California, United States of America.
- O'Neill, I.J., Handal, J., 2021. <https://www.jpl.nasa.gov/news/nasa-analysis-earth-is-safe-from-asteroid-apophis-for-100-plus-years>. NASA Jet Propulsion Laboratory.
- Pavur, J., Martinovic, I., 2020. SOK: Building a Launchpad for Impactful Satellite Cyber-Security Research. eprint arXiv:2010.10872.
- Planck, M., 2009. Air and Space Law, in: *Max Planck Institute for Comparative Public Law and International Law, World Court Digest (Formerly Fontes Iuris Gentium)*. Springer, Berlin, Heidelberg.
- Plotnek, J.J., Slay, J., 2021a. Cyber terrorism: A homogenized taxonomy and definition. *Computers & Security* 102, 102145. <https://doi.org/10.1016/j.cose.2020.102145>
- Plotnek, J.J., Slay, J., 2021b. Satellite Cyber Resilience Whitepaper. SmartSat CRC, Adelaide, Australia.
- Santamarta, R., 2014. SATCOM Terminals: Hacking by Air, Sea, and Land. IOActive, United States of America.
- Schrogl, K.-U., Hays, P.L., Robinson, J., Moura, D., Giannopapa, C. (Eds.), 2020. *Handbook of Space Security*. Springer, New York, United States of America.
- Sheehan, M., 2020. Defining Space Security, in: Schrogl, K.-U., Hays, P.L., Robinson, J., Moura, D., Giannopapa, C. (Eds.), *Handbook of Space Security*. Springer, New York, United States of America, pp. 7–21.
- Steinberger, J.A., 2008. A Survey of Satellite Communications System Vulnerabilities. Joint Electronic Warfare Center, United States.
- Stephens, D., 2021. The Woomera Manual [WWW Document]. URL <https://law.adelaide.edu.au/woomera/> (accessed 3.29.21).
- UCS, 2020. Union of Concerned Scientists Satellite Database [WWW Document]. Union of Concerned Scientists. URL <https://ucsusa.org/resources/satellite-database> (accessed 12.16.20).
- USDoD, 2015. Space Domain Mission Assurance: A Resilience Taxonomy. Office of the Assistant Secretary of Defense for Homeland Defense & Global Security, United States.

- van der Watt, R., Slay, J., 2021. Modification of the Lockheed Martin Cyber Kill Chain (LMCKC) for cyber security breaches concerning Low Earth Orbit (LEO) Satellites. Presented at the 16th International Conference on Cyber Warfare and Security.
- Wheeler, W.A., Cohen, N., Betser, J., Ewart, R.M., 2018. Cyber Resilient Flight Software for Spacecraft, in: AIAA SPACE and Astronautics Forum and Exposition. American Institute of Aeronautics.
- Wright, D., Laura, G., Lisbeth, G., 2005. The Physics of Space Security: A Reference Manual. American Academy of Arts and Sciences, Cambridge, United States of America.
- Zissis, C., 2010. China's Anti-Satellite Test. Council on Foreign Relations.

Assessment of Cybersecurity Risks: Maritime Automated Piloting Process

Jouni Pöyhönen and Martti Lehto

University of Jyväskylä, Finland

jouni.a.poyhonen@jyu.fi

martti.j.lehto@jyu.fi

Abstract: A modern society is a combination of several critical infrastructures, of which international and national maritime transportation systems are essential parts. Digitalization makes it possible to increase levels of autonomy in maritime systems. It also means fully existing cyberenvironments in maritime processes. In cyberenvironments, it is crucial there is trustable information communication between system elements of the process, alongside the usability, reliability, and integrity of systems data in the operating environment. In order to develop maritime autonomy in Finland the Sea4Value / Fairway (S4VF) research program has been developed. At the first stage of the program, the main goal is to create automated fairway piloting feature in the near future. An automated remote piloting process, “ePilotage,” will be a complex system of systems entity. This paper provides a research approach to investigating the cybersecurity risks at the system levels of process. It emphasizes the importation of comprehensive risk assessment to increase the cybersecurity of fairway operations. The findings of the study are located in cybersecurity risks in critical information flows between the main system blocks of the fairway process. The research question is “How can the cybersecurity risks of automated remote fairway operations be evaluated?” The main findings are related to the probabilities of the risks in all levels of process stakeholders’ responsibilities. Risk assessment methodology, that has been described, is based on attack probabilities against probabilities to defend actions of adversarial in use of communication technologies. Risks assessment factors have been identified and the risk assessment tool have been proposed.

Keywords: maritime digitalization, automated piloting, cybersecurity, risks assessment, probability

1. Introduction

A modern society is a combination of several critical infrastructures, of which international and national maritime transportation systems are essential parts. Digitalization and information and communication technologies (ITC) make it possible to increase levels of autonomy in maritime systems: “ICTs are increasingly intertwined in our daily activities. Some of these ICT systems, services, networks and infrastructures form a vital part of the European economy and society, either providing essential goods and services or constituting the underpinning platform of other critical infrastructures” (EU Commission, 2009). This also means fully existing cyberenvironments in maritime processes. Crucial elements in the cyberenvironment include trustable information networks and the usability, reliability, and integrity of systems data in the operating environment.

In order to develop maritime autonomy in the first stage in Finland, the Sea4Value / Fairway (S4VF) research program has been launched to create automated remote pilotage fairway features (ePilotage) in the near future: “As defined in the ePilotage Act, pilotage refers to activities related to the navigation of vessels in which the pilot acts as an advisor to the master of the vessel and as an expert on the local waters and their navigation. The purpose of pilotage is to enhance the safety of vessel traffic and prevent environmental damage generated by vessel traffic” (Finnpilot Pilotage Ltd, 2020).

The ePilotage environment of the S4FV is an example of a system of systems (SoS) in which an increased number of digital solutions are entering new environments where traditional engineering solutions are still in use. This development introduces an increased risk of a malicious cyberadversary taking deliberate actions against the system, which is why the threat analyses should be done according to the principles within the system-of-systems threat model (Bodeu & McCollum, 2018).

At the strategic level of SoS thinking, we present a description of the situation with non-technical terminology. This includes different types of adversaries based on their motivational factors: cybervandalism, cybercrime, cyberespionage, cyberterrorism, cybersabotage, and cyberwarfare. In the case of cybervandalism, the arrival of a controversial vessel in a fairway might trigger actions. The controversy might be with the cargo, the vessel’s operations, or the vessel’s owner. For cybercrime, valuable cargo is more tempting, with financial gain as the motivation. Cyberespionage can include business or political espionage. Political factors may arise from national or international issues. On the national side, hacktivism supporting strikes in harbors could be one scenario. In

the worst case, international tensions in the region could escalate to military cyberoperations against vessel traffic (Kovanen, Pöyhönen & Lehto, 2021b).

On the operational level, we have a description of the situation with business continuity, including information on the attacker's capabilities to attack ICT and industrial control systems (ICS; Kovanen, Pöyhönen & Lehto, 2021a).

On the tactical level, we have more technical information on the threat actor's tactics, techniques, and procedures. On this level, the follow-up operations may include constant threat assessments about how the changes in an area can affect the strategic and operational levels. Threat analyses in this scope should include both user and technical views of the process. The threat probability tree model has been used for ship's cybersecurity assessment and to maintain situation awareness as an example of systems threat analyses and the risk assessment method (Hummelholm, Pöyhönen, Kovanen & Lehto, 2021).

In the ePilotage process, trustable information networks and the usability, reliability, and integrity of systems data in the fairway operating environment mainly depend on reliable information flows at the SoS levels. The cybersecurity research question related to risk assessment in this project is as follows: How can the cybersecurity risks of information flows in automated remote piloting fairway operations be evaluated?

This paper presents a research approach for cybersecurity risk assessment at the system level in the remote ePilotage process through the use of probability estimation principles. The findings of the study are situated in cybersecurity risks in critical information flows between the main system blocks of automated remote pilotage fairway and the main findings are related to the probabilities of the risks in all levels of stakeholders' responsibilities.

The paper is organized as follows: section 2 introduces automated remote pilotage fairway features, section 3 presents the basic elements of the research subject, section 4 describes a risk assessment approach done by us, section 5 discusses stakeholder responsibilities for security risks. Section 6 closes the paper with some conclusions.

2. Automated remote pilotage fairway features

This section is based directly on the Sea4Value / Fairway program plan (DIMECC Oy, 2020), which concerns a smart maritime transport system and its automated remote pilotage fairway navigation. This automated fairway navigation ensures a channel by which the existing vessels and future develop of vessels. It enables one path of the digital supply chain. In maritime navigation operations, the sea areas can be distributed to different approach areas of the ports. In the case of fairway navigation, the information requirements and the amount of information needs are related to the accuracy and reliability of safe and efficient fairway navigation. According to the Sea4Value/Fairway research project plan, the main information items required are as follows:

- movement of the ship, particularly the speed and rate of turn (and accuracy of course keeping) and its predicted path
- dimensions and location of the navigable fairway (particularly the limits of the fairway)
- position of the ship and high accuracy of position fixing (need for redundancy in position fixing)
- maritime aids to navigation
- vessel traffic to be encountered and the location of encounter
- environmental conditions
- safety-related communication
- the intended route to destination
- actual route plan to destination including waypoints (wheel over points and turning radius to be used).

(DIMECC Oy, 2020)

Safe navigation in the fairway process requires the combination of accurate and reliable information along with the competence to use it in decision-making and choose the correct action to be taken. The future ePilotage technical environment vision is presented in Figure 1 according to the Sea4Value/Fairway research project plan. It involves intelligent safety equipment, increased situational awareness and a shore pilotage center (DIMECC Oy, 2020).



Figure 1: Vision of the future ePilotage technical environment (DIMECC Oy, 2020)

It should also be noted that the navigation and pilotage process consist of cooperation and communication among many stakeholders. In the present situation, this is often done by the pilot from the bridge onboard. The relevant stakeholders include:

- law enforcement and maritime authorities
- fairway services ensured by the state (e.g. pilotage and icebreaking)
- commercial fairway and port services (towing, salvage, port, and cargo operations)
- safety authority for weather and sea observations and forecasts.

(DIMECC Oy, 2020)

The list above is not exhaustive and not all stakeholders have a role at all times in the pilotage process. However, it is important to recognize that communication ability must be considered when designing new methods for fairway navigation. This is particularly important with special services, such as towing and icebreaking, as the pilot is often the only person onboard with the competence for using their assistance.

3. Basic elements of cybersecurity for an automated remote piloting fairway

The paper “Basic elements of cybersecurity for an automated remote piloting fairway system” (Pöyhönen, Kovanen & Lehto, 2021) presents a description of a research framework for cybersecurity study of a maritime automated remote pilotage fairway system. It is a combination of a five-layer cyberstructure of an organization (Figure 2) and the maritime security management system (MSMS) with aspects of cybersecurity and the block diagram of an automated remote pilotage system (Figure 3).

In general, an organization’s cybersecurity management requires comprehensive awareness on the system level. The awareness of an organization and decision-makers can be seen as system-level awareness arrangement. It is possible to integrate an organization’s three decision-making levels into a five-layer cyberstructure (Figure 2) in order to have a comprehensive system view of that organization’s cybersecurity environment (Pöyhönen & Lehto, 2020).

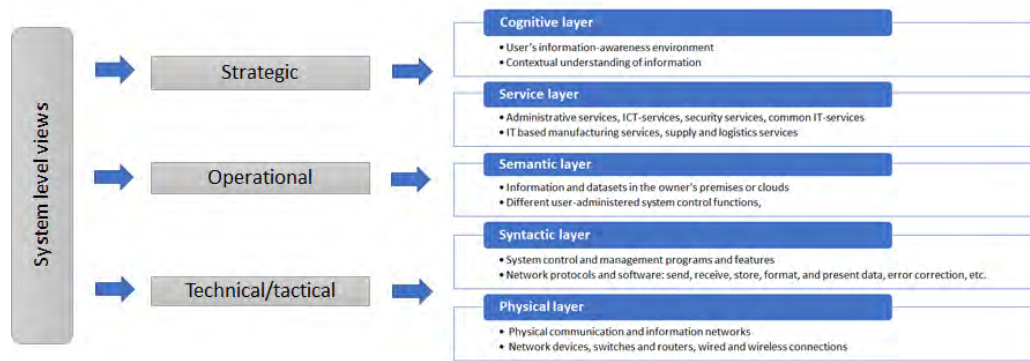


Figure 2: System-level view of organizational cybersecurity (Pöyhönen & Lehto, 2020)

In the paper “The Maritime Security Management System: Perceptions of the International Shipping Community” (2007), Vinh V. Thai and Devinder Grewal from the Australian Maritime College emphasize the importance of identifying key shore-based and near-shore activities associated with maritime security management. The MSMS includes the following components: activities, main players, organizational relationships, security dimensions, security elements, and criteria. The same basic framework has been used in Sea4Value research. The cybersecurity aspects and features in our research are similar to the activities component, meaning the ePilotage process itself, and security elements, which include capabilities such as people, processes, and technology. These can be seen as common elements in cybersecurity management of the process. The other features of cybersecurity management include security dimensions such as those systems that are needed for operations, security awareness, and training. The risk management and continuity enhancement of system operations establish the criteria for security management. In that sense, the strategy, operational and technology/tactical viewpoints on the systems of stakeholders support a holistic approach to security. The rest of cybersecurity system features include stakeholders and their organizational relationships (Pöyhönen et al., 2021).

Figure 3 presents the cybersecurity system sides as a block diagram for the case of the ePilotage process. The direct and main process systems are fairway systems, ship systems and control center systems. The automated remote pilotage operations also need support processes like vessel traffic service (VTS) and weather forecast services. Situation awareness from other vessels and stakeholder’s operation are also essential information and support processes. The complete system is a complex SoS, including information flows between process elements mentioned above with cybersecurity risks (Pöyhönen, et al., 2021).

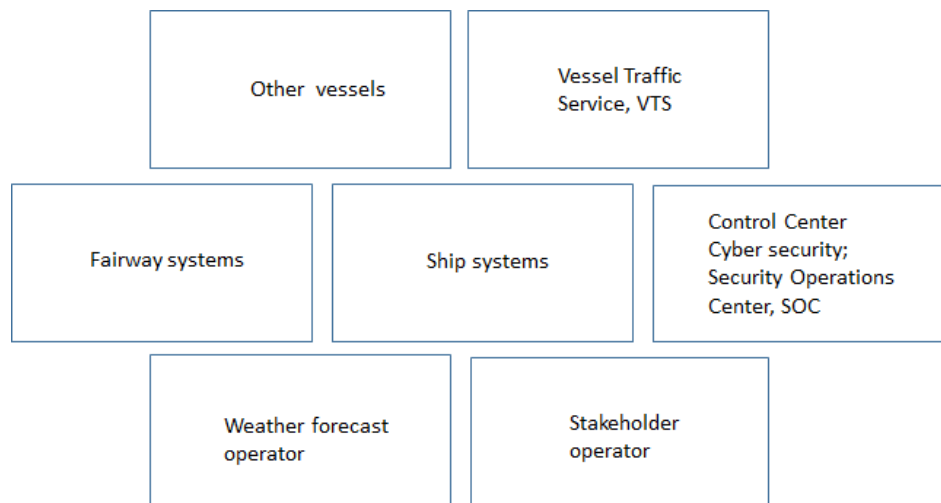


Figure 3: Block diagram of an automated remote pilotage system (Amro A., 2019, modified; Pöyhönen, et al., 2021)

4. Risk assessment

The development of ePilotage fairway systems increasingly use ICT systems and ICS systems to exchange information between systems on the navigation process. It happens via different data and communication lines. These information flows use legacy as well as new technologies, which are very important parts of cybersecurity risk analysis work. The recognition of these technologies enables us to identify different functions at the system level, carry out risk assessments and identify their residual risks with sufficient accuracy. In the same way, the dependencies of different information systems need to be considered and, based on these dependencies, security and cybersecurity risks need to be identified. In this chapter we present a probability approach to cyberattacks versus a probability to defend attacks and at the end to evaluate cybersecurity risks related to the information flows of ePilotage operations. We do argue that many previously described detailed calculation principles of cyber-risks, such as the paper “Dynamic Security Risk Management Using Bayesian Attack Graphs” (Poolsappasit, Dewri & Ray, 2012), are too complicated to use in daily life. We thought that there should be a more practical method and tool to make cyber-risk assessments in SoS environments.

A cyber-threat model captures information about potential cyberthreats against a system, an enterprise, an SoS, a region, or a critical infrastructure (CI) sector. A cyber-threat model can serve as a basis for a variety of tasks in different scopes. Comprehensive cybersecurity needs wide scale of analysis of a system of systems (or sub-system) against a set of threat events. It can be often impractical and, in that sense, analysis of system of systems could rely on the development and use of threat scenarios. A threat scenario could include the picture of a potential threat and the result of harmful consequences (Bodeu & McCollum, 2018, 1).

In order to have a practical SoS threat scenario, an adversary can operate by targeting tailored attacks or attacks that reflect the broader scale of the system. Rather than focus only on specific assets or asset types, SoS threat modeling should also focus on sub-sectors or specific business functions. For SoS, three aspects of the scope can be considered (Bodeu & McCollum, 2018, 13):

- Institutional scope: whether the adversary’s targeting focuses on a single institution or on a family of institutions.
- Functional scope: how narrowly or broadly a business function, sub-sector, or sector is targeted.
- Technical scope: how specifically or broadly are technologies (or specific systems) targeted.

At the strategic level (institutional scope), there must be a description of the situation with non-technical terminology. This includes, for example, discussion of the attacker’s motivations by studying different types of adversaries, such as cybercrime or terrorism, etc. At the operational level (Functional scope), there must have a description of the situation with business continuity. This includes, for example, information of the attacker’s capabilities to attack against ICT and ICS systems. At the technical/tactical level (Technical scope), there must be more technical information on the threat actor’s tactics, techniques, and procedures.

The Bayesian attack graph model have been studied in several areas of security risk management. The paper “Dynamic Security Risk Management Using Bayesian Attack Graphs” (Poolsappasit et al., 2012) proposes a risk management framework using Bayesian networks in order to quantify the chances of network compromise at various levels of system constructions. In the same sense, various threat risk analysis schemes have been developed to recognize the attack and implement the security safeguards to protect the ICT system asset from cyber-attacks (Wang & Liu, 2014): “Attack trees (AT) technique plays an important role to investigate the threat analysis problem to known cyber-attacks for risk assessment”. An attack graph is based on a probabilistic metric model and can be used to quantify the cybersecurity issues of an SoS environment.

In this paper, an attack tree graph is used to represent the relationship between threat and defense actions in the ePilotage process. At the SoS level, it is more than a metric model a way of thinking, because there are many layers in the system configuration. Exact probability calculation is therefore complicated, and results can be inaccurate. According to our experiences, as the authors of this paper and other ePilotage research group members, we are familiar with the attack tree graph as a tool for risk assessment. The result of this is representing the likelihood of an attacks against the likelihood of defense against such an attacks. The final probability of success of defense measures versus attacks will be estimated and the most serious attacks will be recognized and prioritized. This probability evaluation work is proposed to be done by cybersecurity experts by utilizing all relevant information that is available from the cybersecurity features of the used technology in the

current information flow (see Table 3) and as well as information from stakeholders' capabilities to have defense measures. In this sense we would like to propose the use of Delphi method principle to make relevant threat analysis and risk level estimations from the systems. It is a useful way of thinking about likelihood and probabilities at the system level of a process or an organization.

The risks probability estimation can be extended to the system level as described below. The National Institute of Standards and Technology (NIST) released recommendations as "Framework for Improving Critical Infrastructure Cybersecurity" (2018) for owners and operators of critical infrastructure to help them identify, assess, and manage cyber-risks. The Framework Core part of the guidance has a set of cybersecurity activities, outcomes, and informative references that are common across sectors and critical infrastructure. Elements of the Framework Core provide detailed guidance to help an organization align and prioritize its cybersecurity activities with its business/mission requirements, risk tolerances, and resources. The Framework Core consists of five concurrent and continuous Functions: Identify, Protect, Detect, Respond, Recover. There is mentioned that "The Framework is not a one-size-fits-all approach to managing cybersecurity risk for critical infrastructure" (NIST, 2018).

In this paper, functions according to NIST are proposed to exploit a system risk framework. The meaning of its notations are described in Table 1. Cybersecurity risk assessment can be set to S4F ePilotage process by investigating probabilities and using the elements of the table. In other words, the table can be used as a risk assessment tool by investigating the probabilities of each element of it. After that, it is practical to apply probability tree principles. In Figure 4, the probability tree is described as using Defense probability P_D against Attack probability P_A in the ePilotage process. Cyberattacks (A) in this Sea4Value ePilotage process are the same as the "Attack Identification" and located on all levels of system responsibilities. The P_A attack probability (P_{SOT}) to defend attack probability P_D (P_P, P_D, P_M, P_R) is related to the combination of cybersecurity capabilities (people, processes, and technologies) (Jacobs, von Solms & Grobler, 2016), using "Protection" (P), "Detection" (D), "Countermeasure" (M) and "Recovery" activities according to Table 1. The entire risk assessment process needs to have experienced cybersecurity professionals performing this work.

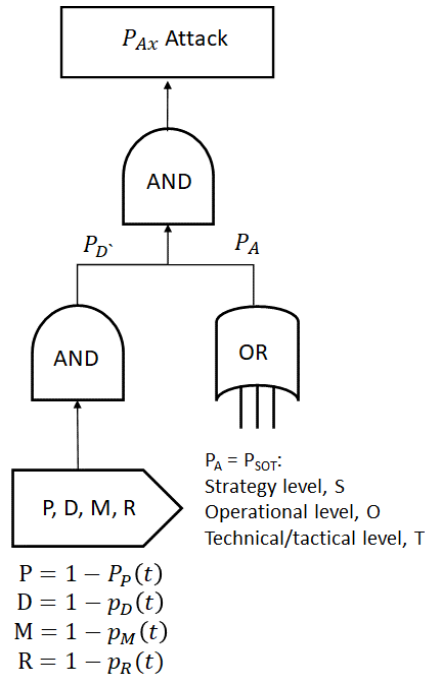


Figure 4: Probability tree; Defense probability P_D against Attack probability P_A in the ePilotage process (Wang & Liu, 2014, mod.)

Table 1: Meaning of Notations of ships risk probabilities (NIST, 2018; Hummelholm et al., 2021)

ACTION	EXAMPLES	NOTATION
ATTACK IDENTIFICATION	1. Attacks at strategy level (S) 2. Attacks at operational level (O) 3. Attacks at the technical/tactical level (T)	A
PROTECTION CATEGORIES	1. Identity Management and Access Control 2. Awareness and Training 3. Data Security 4. Information Protection Processes and Procedures, 5. Maintenance 6. Protective Technology (Port scan, FIREWALL, IDS, IPS, SIEM...)	P
DETECTION CATEGORIES	1. Anomalies and Events 2. Security Continuous Monitoring and Detection a) SOC	D
COUNTERMEASURE (RESPOND) CATEGORIES	1. Conducting Response Planning 2. Communications and Analysis: a) Real time Situation Awareness b) OODA-procedure 3. Mitigation and Improvements	M
RECOVERY CATEGORIES	1. Recovery Planning 2. Improvements 3. Communications.	R

The probabilistic success of attacks, $P(t)$, against the defense of system x can now be evaluated and calculated as follows adapting the principle in “Threat Analysis of Cyber-Attacks with Attack Tree+” (Wang & Liu, 2014, mod.)

$$P_{Ax}(t) = P_A P_D = (P_{SOT})(1 - P_P(t))(1 - p_D(t))(1 - p_M(t))(1 - p_R(t)) \quad (1)$$

where, as a function of time t , a successful attack against a system x , P_{Ax} has attack success probabilities P_A (P_{SOT} : strategy level S ; operational level, O ; technical/tactical level, T) reduced by a defending mechanism, P_D : protection P , detection D , countermeasure M and recovery R having the respective success probabilities P_P , P_D , P_M and P_R .

Table 2 illustrates the cybersecurity dimensions of systems of the ePilotage process. (Pöyhönen et al., 2021). Table 2 also includes subsystems for information flows and the relevant technologies in the columns. The information technologies behind the numbers are explained in Table 3. The risks assessment process should be based on knowledge of threats, the vulnerabilities of these technologies and the capabilities of relevant stakeholders to recognize and defend the attacks behind the threats.

Table 2: ePilotage system dimensions, an example of subsystem and ICT technologies

ePilotage subsystem	Ship	Other Ships	Control Center	VTS	Weather Forecast	Stakeholders	Fairway
AIS	2	2					
Telecomm.	3	3	3	3	3	3	
Weather	1	1	7		7		
Radar	1	1	7				
GNSS	4	4					
Internet	3	3	7	7	7	7	
Cloud	3	3	7	7	7	7	
Drone	3,4,5,6	3,4,5,6					
Fairway sensors	5	5					5

In general, organizations should identify, characterize, and provide representative examples of attacks, and likelihood determinations promote a common terminology and frame of reference for comparing and addressing risks across disparate mission/business areas. Organizations can also select appropriate risk assessment methodologies, depending on organizational governance, culture, and how divergent the missions/business functions are within the respective organizations (NIST 800-39, 2011).

It is proposed that in the ePilotage process, the risk assessment methodology will be based on attack probabilities against probabilities to defend actions of adversarial in communication technologies and thus in information flows between the main system of systems. Table 3 presents a tool to recognize risks by using relevant references. It is also a tool for risk estimation against information flows. Risk assessments can be conducted at any of the process decision levels (Strategy level, S, Operational level, O, Technical/tactical level, T) with different objectives and the utility of the information produced from people, processes, and technology. At the end of the risk assessment work, total risks can be addressed to the security dimensions of the ePilotage process by using Table 2.

Table 3: Risk assessment of information flows

Cybersecurity risks/ ICT technology	Cybersecurity information: References	Risk evaluation $P = \text{act, mot, vul}$ $D = \text{defense}$
1. VHF/(HF)	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
2. Satellite	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
3. GSM/4G/5G	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
4. GNSS/GPS	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
5. MESH	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
6. WiFi	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$
7. Capel network	Cyberthreats CS feature Situation awareness	$P_A = f(S, O, T)$ $D = f(P, D, M, R)$

After risk assessment process, ePilotage stakeholders can respond to risks in a variety of ways. These could include risk acceptance, risk avoidance, risk mitigation, risk sharing, risk transfer, or a combination of the above (NIST 800-39, 2011). The stakeholders should monitor ePilotage risks of operations continuously including the purpose, type, and frequency of monitoring activities. Risk monitoring provides stakeholders with the means to verify compliance, determine the ongoing effectiveness of risk response measures, and identify risk-impacting changes to information systems and environments of operation (NIST 800-39, 2011). Analyzing monitoring results allows the stakeholders to maintain situations awareness (SA) by using the risk being incurred, highlight the need to revisit the risk management process, and initiate process improvement activities as needed and communicate with other stakeholders.

5. Discussion

In addition to the attack probability in information flows of the ePilotage process, it is necessary to consider the cybersecurity procedures of ships, every stakeholder and fairway service producer. Threats to information and communication systems and as well as to industrial control systems can include purposeful attacks, vulnerabilities in the systems, and human or machine errors, causing great harm to the services and economy of maritime traffic. Therefore, it is imperative that all stakeholders (Table 2: responsible security dimensions) and at all levels of decision processes understand their responsibilities and are held accountable for managing information security risks. The cybersecurity architecture of stakeholders is an integral part of a comprehensive cybersecurity architecture in any process case (Pöyhönen & Lehto, 2020). It represents system resilience along with providing cybersecurity capabilities for maritime traffic and its operation continuity resilience.

In order to make the piloting process cybersecure, the risks for the main information flows at the SoS level should be investigated carefully. The proposed risk assessment is an efficient method and key element to make threats and defense capabilities visible. Cybersecurity is an essential part of the trust process. The main advantage of

the proposed approach is to achieve good results in near real-time effective attack modelling and security evaluation by using constant awareness of threats, vulnerabilities, and defense procedures.

In that sense, all stakeholders should have real-time situation awareness (SA) of the ePilotage process. In addition, they should use an Observe – Orient – Decide – Act (OODA) loop for SA information sharing between each other (Pöyhönen, Rajamäki, Ruoslahti & Lehto, 2020). These are the key features for conducting response procedures to the risk management across ePilotage process.

6. Conclusion

In order to develop maritime autonomy in the first stage in Finland, the Sea4Value / Fairway (S4VF) research program has been launched to create automated remote fairway pilotage features in the near future. ePilotage, the automated remote pilotage system, is a process and an essential part of the critical maritime traffic and transportation supply chain. The fairway and its stakeholders' ICT and ICS systems are together a complex SoS entity, characterized by a conglomeration of interconnected networks and operational dependencies. This new research program increases the amount digital solutions, stakeholders, and subcontractors in maritime fairways. However, there will also be a continuing need for traditional engineering solutions for a long time to come. This development introduces increased risks of a malicious cyberadversary taking deliberate actions against the systems. A comprehensive risk assessment of the ePilotage process should be done according to the principles of SoS modeling.

This paper has established a research framework for the cybersecurity risk assessment of maritime automated remote ePilotage fairway systems and processes. The framework uses probability evaluation in main ICT information flows between the main fairway systems. The risk assessment methodology that has been used is based on attack probabilities against the probabilities to defend against adversarial actions in the use of communication technologies. Risks assessment factors have been identified and the risk assessment tool has been described. It is a way of thinking about risks and risk prioritization. These are needed to answer the research question: "How can the cybersecurity risks of automated remote piloting fairway operations be evaluated?"

Protecting the ePilotage system against cyberthreats implies measures taken based on risk assessment, and they ensure confidentiality, integrity, and the availability of primarily digital information in the operating processes being examined. The measures should be highly significant for the overall availability of the systems that support the stakeholders' processes in the ePilotage environment. Operational availability plays a key role in achieving operational continuity and promoting the reliability of activities. Cybersecurity risk management and information security are mandatory features from the point of view of operational trust, continuity, reliability, and resiliency.

References

- Amro, A. (2019). Communication and Cybersecurity for Autonomous Passenger Ferry (Autoferry).
- Bodeu, D. J. & McCollum, C. D. (2018). System-of-systems threat model. The Homeland Security Systems Engineering and Development Institute (HSSEDI) MITRE: Bedford, MA, USA
- DIMECC Oy, 2020. Sea4Value / Fairway program (S4VF).
- Finnpilot Pilotage Ltd, (2020). <https://finnpilot.fi/en/pilotage/what-is-pilotage/>
- EU Commission, 2009. Critical information infrastructure protection (2009), COM (2009) 149 final, Commission of the European Communities, Brussels, 30.3.2009.
- Hummelholm, A., Pöyhönen, J., Kovanen, T. & Lehto, M. (2021). Cyber Security Analysis for Ships in Remote Pilotage Environment. Submitted to be published in ECCWS 2021 - 20th European Conference on Cyber Warfare and Security. 24th - 25th June 2021, Chester, UK.
- Jacobs, P. C., von Solms, S. H. & Grobler, M. M. (2016). Towards a framework for the development of business cybersecurity capabilities. International Conference on Business and Cyber Security (ICBCS), London, UK. The Business and Management Review, Volume 7 Number 4, 51–61.
- Kovanen, T., Pöyhönen, J. & Lehto M. (2021a). ePilotage System of Systems' Cyber Threat Impact Evaluation. Proceedings of the 16th International Conference on Cyber Warfare and Security ICCWS 2021. p. 144-151.
- Kovanen, T., Pöyhönen J. & Lehto M., (2021b). Cyber Threat Analysis in the Remote Pilotage System. Submitted to be published in ECCWS 2021 - 20th European Conference on Cyber Warfare and Security. 24th - 25th June 2021, Chester, UK.
- National Institute of Standards and Technology, NIST 800-39, (2011). Managing Information Security Risk. Organization, Mission, and Information System View. U.S. Department of Commerce
- National Institute of Standards and Technology, NIST. (2018). Framework for Improving Critical Infrastructure Cybersecurity, April 16, 2018

- Poolsappasit, N., Dewri, R. & Ray, I. (2012). Dynamic Security Risk Management Using Bayesian Attack Graphs. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 9, NO. 1, JANUARY/FEBRUARY 2012
- Pöyhönen, J., Rajamäki, J., Ruoslahti, H. and Lehto, M. (2020). Cyber Situational Awareness in Critical Infrastructure Protection. Article approved 2nd March 2020 to Cyber Security of Critical Infrastructure 2020 (CYSEC2020) conference, October 27th, 2020 - October 28th, 2020. Dubrovnik. Croatia.
- Pöyhönen J. & Lehto M. (2020). Cyber security: Trust based architecture in the management of an organization security. The 18th European Conference on Cyber Warfare and Security ECCWS2020, 25-26 June 2020, University of Chester, UK, pages 304-313
- Pöyhönen J., Kovanen T. & Lehto M., (2021). Basic elements of cyber security for an automated remote piloting fairway system. Proceedings of the 16th International Conference on Cyber Warfare and Security ICCWS 2021. p. 299-308.
- Thai, V.V & Grewal, D. (2007). The Maritime Security Management System: Perceptions of the International Shipping Community. Article in Maritime Economics & Logistics · June 2007.
- Wang, P. & Liu, J. C. (2014). Threat analysis of cyber-attacks with attack tree+. Journal of Information Hiding and Multimedia Signal Processing, 5(4).

Utilizing Switch Port Link State to Detect Rogue Switches

Travis Quitiquit and. Vijay Bhuse

Grand Valley State University, Allendale, MI, USA

quitiquit@mail.gvsu.edu

bhusevij@gvsu.edu

Abstract: There are many methods to detect rogue wireless access points, but the same case can not be said for rouge switches on a LAN network. Detecting these rogue switches is key to security of any organization. The introduction of a rogue unmanaged network switch has the potential to cripple a network. These types of switches pose a big risk because they are usually plug and play types of devices and can prove difficult to track. A switch becomes rouge when it is connected to a network without proper authorization. Rogue switches are a huge threat to the security and reliability of any network. An attacker could use a rogue switch to launch an attack or spy on network traffic information. Many organizations these days implement a “bring your own device” policy that can prove to be a daunting task to monitor for any network administrator. It is important that these rogue network switches are not introduced to a network, whether by accident or in a malicious attempt. The vulnerability that is introduced could comprise the confidentiality of network messages, degrade network performance, or even allow hackers or authorized users access to critical network infrastructure and data. In this paper we present a method that can help detect these rogue switches by monitoring the ethernet frames across the network and looking at the link state of the network switch ports. We will be using Wireshark, a Windows Computer, and a local switch setup to test methods for detecting a rogue switch. In our scenario we were able to provide some evidence of a method that could be used in conjuncture with other rules and policies to detect rogue switches connected to a network. We were able to determine based on the port link state that there was another device, most likely a rogue switch between the good switch and the computer.

Keywords: rogue switch, link state, port monitoring, network security, LAN, network detection

1. Introduction

There are two main types of switches that can be introduced to a network, managed and unmanaged. A managed switch allows for greater control over the settings of the switch and can prioritize network traffic. These are typically not the rouge switches you think about since managed switches take more effort to configure to work on a network. On the other hand, the unmanaged switches are what pose the biggest risk to an organization. These are essentially plug and play devices with no settings to configure (Services, 2021). These unmanaged switches can pose a large risk to any network because they do not have an IP address, and most do not have MAC addresses for device tracing.

Using managed switches in your organization, you should be able to monitor all traffic on the network for MAC addresses among other things. You can then enforce rules and policies such as only one MAC address can be routed to each port on the managed switch. If a rouge switch was used to plug into an existing port, you would notice this because there would be more than one MAC addresses routed to a given port.

A switch becomes rouge when it is connected to a network without proper authorization. Rogue switches are a huge threat to the security and reliability of any network. An attacker could use a rogue switch to launch a cyber-attack or spy on network traffic information. Many organizations these days implement a bring your own device policy that can prove to be a daunting task to monitor for any network administrator. It is important that these rogue network switches are not introduced to a network, whether by accident or in a malicious attempt. The vulnerability that is introduced could comprise the confidentiality of network messages, degrade network performance, or even allow hackers or authorized users access to critical network infrastructure and data (Bhuse, 2019).

We live in an IP connected world and internet protocol has become the standard for all connected technology. The dilemma today is the greater connectivity the greater challenges if you’re an IT or security manager. Network growth and complexity is skyrocketing. Before the explosion of connected devices, networks usually grew with the number of connected users. These days networks must scale three to five times for each employee based on the number of IP devices we use to do our work. This means the challenge of managing and securing endpoints has also become much more difficult (Solarwinds, 2019).

When you consider all the connected devices on a network: PCs and laptops, VOIP phones and BYOD devices, tablets, smartphones. There is also another type of device lurking out there and those are rogue devices that have the potential to cause harm to your network. The larger the network, the greater the potential risk. All it takes is one rogue device to inflict chaos on a network and organization (Solarwinds, 2019).

The goal of security is to implement security controls at the network layer. The network layer represents the logical area of vulnerability and point of intrusion. Some goals of network security at this level should include identifying authorized devices, building network access controls, and monitoring network connected devices. The network switches represent the best touch point for implementing many of these goals, because network switches are the part of infrastructure which is closest to network cabling. Another good fact is that switches operate at OSI layers 1 and 2, which can detect activity on your network which is not detectable at higher levels (3-7) (Solarwinds, 2019).

Using some methods of security that are already in place today (Bhuse, 2019) along with the test methods of this paper we came up with an experiment to see if detecting rogue switches utilizing port link state is beneficial. Some setup and assumptions of the experiment are that the good switches are managed switches with a rule that each port can only have one MAC address. It is also assumed that the rogue switch is plugged into an enterprise network where we have access to the good network switches and to a computer that is connected to the rogue switch that we have access to get logs from.

2. Related work

The job of detecting a rogue network switch can be a difficult task because most unmanaged switches do not have IP addresses. They may also lack the full network protocol stack and finding the true location of an unmanaged switch can prove difficult even if it is detected (Bhuse, 2019). There are some solutions out there that help in detecting rogue switches. One method that has been investigated is leveraging the Dynamic Host Control Protocol (DHCP). In this method you can use the DHCP request to see if a computer requests an IP if the connection from the workstation to the network switch is severed and it doesn't request an IP after reconnecting the network. Another useful technique that can be deployed today to help eliminate rogue network switches is implementing MAC address whitelisting. An organization could create a whitelist of valid MAC addresses of known switches, and the valid switches would know of its network's valid MACs to communicate with. If a switch received a packet from a MAC address that is on the whitelist, it would forward that packet and add that MAC to its forwarding table. If a switch received a packet from an unknown MAC that was not on the whitelist, it could flag the packet as potentially originating from a rogue device (Bhuse, 2019).

2.1 Detecting rogue access points in a computer network

Keohane came up with a computer implemented method, data processing system, and computer program product for discovering an unauthorized router in a network. The process first obtains a physical address of a suspected router or destination device. A data packet is created which contains a destination MAC field, a destination internet protocol field, and a time to live field. The destination MAC field includes the physical address of the destination device, where the destination internet protocol field includes a fake internet protocol address, and where the time to live field includes a value indicating the data packet has exceeded a time limit. The data packet is sent to the destination device using the physical address in the destination MAC field. If a time exceeded message is received from the destination device, the destination device is determined to be enabled for routing (Keohane, 2011).

2.2 Rogue router hunter

Lin and Chen produced a method of detecting an unauthorized wireless access point in a network using a detector. A rogue access point detector receives an incoming data packet which is scanned for a time expiration value. The time expiration value may be a time to live value. It is determined whether the time expiration value is the same as a threshold time expiration value. If the time expiration value is not the same as the threshold value, it is determined whether the incoming data packet was routed through an authorized access point in the network. If it is determined that the packet is not being routed from an authorized access point, a security alert is sent to the network, which could be a device like an administrators' computer (Lin, 2011)

3. Problem

The main question today is how do you detect unmanaged switches on a network? They don't have IP addresses or any other way to be contacted directly. The typical personal or desk switch may have four or five ports, a power plug, and LED status lights. There is no direct connection to them like a serial cable to interface with.

This perked my interest as a potentially interesting problem to solve. If you had the time and money, you could just brute force the problem by picking one of your own (managed) switches and tracing out from each port to the rogue system. If you encounter another device along the way, then you have your proof of a rogue switch.

This doesn't seem feasible for large enterprises and doesn't seem to be time or cost effective. That sort of approach does not scale. It might be good if you suspect a certain person or know of an exact location of a breach, but it would be difficult in a large implementation.

You could also try to monitor the port for traffic for rules or policies such as more than one MAC address on a port. This also takes a fair amount of work, assuming your switch supports getting the traffic information on that port. If your switch supports port security and MAC filtering, you might be able to do that, but again, that requires maintenance, since network changes occur daily.

After pondering this one for a while, we came up with the idea of looking at port link state of the switch and computer. We are assuming that we have control of the switches. What this enables us to do is to enable and disable ports on the switch either manually or through software (if the switch has that capability). We would be able to temporarily turn off a port if we suspected something isn't normal in the condition of the port or it isn't meeting a security rule or policy.

4. Scenario setup

To setup and test our link state theories we used a combination of Windows Computers, Wireshark, a local good network switch, a rogue switch, and a router with DHCP (Netgear, 2021). By using a local switch, we could control and simulate the link state of a port and whether it was on or off. When we toggled the control of the port, we could see the difference in communication packets in Wireshark (Wireshark, 2021). In our scenario testing we made a couple assumptions: we had access to the good network switch, and we had access to a computer that was connected to the rogue network switch. There is a router, which is connected to the good switch, which is connected to a rogue switch, which has a computer connected to it (Figure 1).

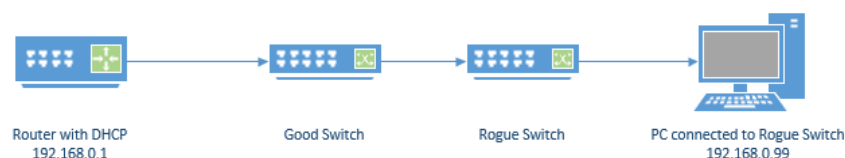


Figure 1: Scenario Setup

5. Results

The scenario was started by changing the state of the downlink port from the good switch to rogue switch from on to off (Figure 2 and Figure 3), this would simulate the port looking just like it had been unplugged from the good switch or turned off through software. The computer we have control of and have access to is connected to the rogue switch where the uplink port was turned off. The computer did not lose its connection because it was still connected locally to the rogue switch. It still saw the active ethernet connection to the switch, even though the uplink on the rogue switch had been severed. The computer was able to log the change in network connectivity which resulted from the good network switch port being deactivated. The rogue switch is now disconnected from the rest of the production network from a logical perspective, but at the physical layer, the computer can still see activity. To find this fact we used the Windows Event Viewer (Figure 4).



Figure 2: Port 4 is computer, Port 8 is uplink



Figure 3: Port 4 is showing activity even though the uplink isn't connected (simulated as disabled)

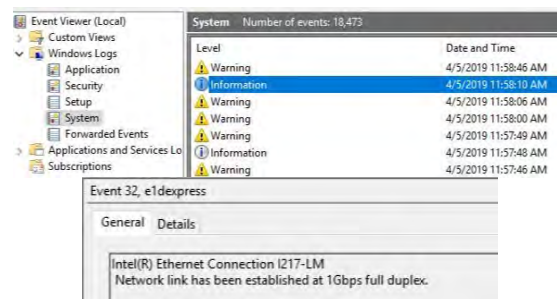


Figure 4: Computer Connected to Rogue Switch Port with uplink disabled

Another useful tool to use during this port enable and disable scenario was Wireshark and Windows command using ARP. We were able to use 'arp -a' and 'arp -d' to refresh the arp table on the local PC. Showing in a normal state you can see the router in the arp table (Figure 5). This screenshot was taken with an active uplink connection on the rogue switch. We could see the router because we had an uplink connection to the good switch which was connected to the router.

When we severed the uplink to the rogue switch, we were able to use the Windows command and clear the arp table. When we ran the command arp -a we found a different result than we got earlier with ethernet connectivity to the rogue switch. You can now see that we're no longer able to see the router in the arp table (Figure 6). This shows that even though our ethernet connectivity to the PC remains in place, we lose routing due to the uplink being disabled.

We were able to find similar findings by using Wireshark and searching for ARP packets. When the uplink on the rogue switch was enabled, we could communicate with the router at 192.168.0.1 (Figure 7). With the uplink disabled we saw a different result in the ARP packets. The computer was asking for router information but couldn't find the information since the uplink was disabled (Figure 8).

```
C:\WINDOWS\system32>arp -a

Interface: 192.168.0.99 --- 0x12
Internet Address      Physical Address      Type
192.168.0.1           40-5d-82-d2-7c-7a    dynamic
192.168.0.255         ff-ff-ff-ff-ff-ff    static
224.0.0.2             01-00-5e-00-00-02    static
224.0.0.251           01-00-5e-00-00-fb    static
224.0.0.252           01-00-5e-00-00-fc    static
239.255.255.250       01-00-5e-7f-ff-fa    static
255.255.255.255       ff-ff-ff-ff-ff-ff    static
```

Figure 5: Normal ARP table with ethernet connectivity

```
C:\WINDOWS\system32>arp -a
```

Interface: 192.168.0.99	---	0x12	
Internet Address		Physical Address	Type
192.168.0.255		ff-ff-ff-ff-ff-ff	static
224.0.0.2		01-00-5e-00-00-02	static
255.255.255.255		ff-ff-ff-ff-ff-ff	static

Figure 6: ARP table with uplink disabled on rogue switch

No.	Time	Source	Destination	Protocol	Length	Info
36	15.871387	Netgear_d2:7c:7a	Lcfcfehe_f0:71:b9	ARP	60	Who has 192.168.0.99? Tell 192.168.0.1
37	15.871413	Lcfcfehe_f0:71:b9	Netgear_d2:7c:7a	ARP	142	192.168.0.99 is at 54:el:ad:f0:71:b9
968	34.891518	Netgear_d2:7c:7a	Broadcast	ARP	60	Who has 192.168.0.62? Tell 192.168.0.1
1045	36.891393	Netgear_d2:7c:7a	Lcfcfehe_f0:71:b9	ARP	60	Who has 192.168.0.99? Tell 192.168.0.1
1046	36.891426	Lcfcfehe_f0:71:b9	Netgear_d2:7c:7a	ARP	142	192.168.0.99 is at 54:el:ad:f0:71:b9

Figure 7: Wireshark ARP packet with uplink enabled on rogue switch

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
4	1.064428	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
5	2.000584	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
7	3.000336	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
11	5.064997	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
12	6.000395	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
14	7.000738	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
18	9.029787	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
19	10.000538	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
21	18.999879	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95
23	22.031067	Lcfefefe_f0:71:b9	Broadcast	ARP	42	who has 192.168.0.1? Tell 192.168.0.95

Figure 8: Wireshark ARP packet with uplink disabled on rogue switch

The network connection shows from the logs and constant ping that the network connection to the computer was always up (Figure 9). The computer never lost its original DHCP reservation from when the uplink was enabled. In the scenario the only time the Windows Event Viewer showed a disconnect from the ethernet connection was when we physically unplugged the ethernet cable from the computer itself (Figure 10).

```
C:\WINDOWS\system32>ping 192.168.0.99 -t
```

```
Pinging 192.168.0.99 with 32 bytes of data:
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Reply from 192.168.0.99: bytes=32 time<1ms TTL=128
```

```
Ping statistics for 192.168.0.99:
```

```
Packets: Sent = 14, Received = 14, Lost = 0 (0% loss)
```

```
Approximate round trip times in milliseconds:
```

```
Minimum = 0ms, Maximum = 0ms, Average = 0ms
```

Figure 9: Ping was consistent even with uplink disabled

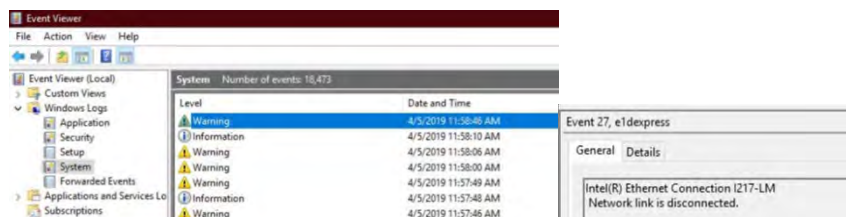


Figure 10: Computer with Ethernet Cable unplugged from rogue switch

Another useful command line tool was used to look at the difference in port and connection status of the link to the computer. Using 'netstat -a' we can display active TCP connections, TCP connections with the listening state, as well as UDP ports that are being listened to. We ran netstat with an active connection to the router (Figure 11) and another test with the uplink disabled and unable to reach the router (Figure 12).


```
C:\WINDOWS\system32\netstat -a
```

Active Connections

Proto	Local Address	Foreign Address	State
TCP	0.0.0.0:135	X270:0	LISTENING
TCP	0.0.0.0:445	X270:0	LISTENING
TCP	0.0.0.0:902	X270:0	LISTENING
TCP	0.0.0.0:912	X270:0	LISTENING
TCP	0.0.0.0:5040	X270:0	LISTENING
TCP	0.0.0.0:5357	X270:0	LISTENING
TCP	0.0.0.0:17500	X270:0	LISTENING
TCP	0.0.0.0:49664	X270:0	LISTENING
TCP	0.0.0.0:49665	X270:0	LISTENING
TCP	0.0.0.0:49666	X270:0	LISTENING
TCP	0.0.0.0:49667	X270:0	LISTENING
TCP	0.0.0.0:49668	X270:0	LISTENING
TCP	0.0.0.0:49669	X270:0	LISTENING
TCP	127.0.0.1:943	X270:0	LISTENING
TCP	127.0.0.1:5939	X270:0	LISTENING
TCP	127.0.0.1:5939	view-localhost:63638	ESTABLISHED
TCP	127.0.0.1:17600	X270:0	LISTENING
TCP	127.0.0.1:50525	view-localhost:50526	ESTABLISHED
TCP	127.0.0.1:50526	view-localhost:50525	ESTABLISHED
TCP	127.0.0.1:63638	view-localhost:5939	ESTABLISHED
TCP	127.0.0.1:63655	view-localhost:63656	ESTABLISHED
TCP	127.0.0.1:63656	view-localhost:63655	ESTABLISHED
TCP	127.0.0.1:63692	view-localhost:63693	ESTABLISHED
TCP	127.0.0.1:63693	view-localhost:63692	ESTABLISHED
TCP	127.0.0.1:63698	view-localhost:63699	ESTABLISHED
TCP	127.0.0.1:63699	view-localhost:63698	ESTABLISHED
TCP	127.0.0.1:63703	view-localhost:63704	ESTABLISHED
TCP	127.0.0.1:63704	view-localhost:63703	ESTABLISHED
TCP	192.168.0.99:139	X270:0	LISTENING
TCP	192.168.0.99:64009	162.125.34.6:https	CLOSE_WAIT
TCP	192.168.0.99:64128	204.79.197.222:https	ESTABLISHED
TCP	192.168.0.99:64137	13.107.42.254:https	ESTABLISHED
TCP	192.168.0.99:64138	131.253.33.254:https	ESTABLISHED
TCP	192.168.0.99:64139	204.79.197.254:https	ESTABLISHED
TCP	192.168.0.99:64142	vpn:https	ESTABLISHED
TCP	192.168.0.99:64143	vpn:https	ESTABLISHED
TCP	192.168.0.99:64157	162.125.18.133:https	ESTABLISHED
TCP	192.168.0.99:64159	52.173.26.181:https	ESTABLISHED
TCP	192.168.0.99:64160	13.107.5.88:https	ESTABLISHED
TCP	192.168.0.99:64161	1drv:https	ESTABLISHED
TCP	192.168.0.99:64163	52.168.20.22:https	TIME_WAIT
TCP	192.168.0.99:64164	52.165.171.165:https	ESTABLISHED
TCP	192.168.0.99:64165	52.114.6.46:https	ESTABLISHED
TCP	192.168.0.99:64166	52.114.6.46:https	ESTABLISHED
TCP	[::]:135	X270:0	LISTENING
TCP	[::]:445	X270:0	LISTENING
TCP	[::]:5357	X270:0	LISTENING
TCP	[::]:17500	X270:0	LISTENING
TCP	[::]:49664	X270:0	LISTENING
TCP	[::]:49665	X270:0	LISTENING
TCP	[::]:49666	X270:0	LISTENING
TCP	[::]:49667	X270:0	LISTENING
TCP	[::]:49668	X270:0	LISTENING
TCP	[::]:49669	X270:0	LISTENING

Figure 11: Netstat with uplink connected

```
C:\WINDOWS\system32\netstat -a
```

Active Connections

Proto	Local Address	Foreign Address	State
TCP	0.0.0.0:135	X270:0	LISTENING
TCP	0.0.0.0:445	X270:0	LISTENING
TCP	0.0.0.0:902	X270:0	LISTENING
TCP	0.0.0.0:912	X270:0	LISTENING
TCP	0.0.0.0:5040	X270:0	LISTENING
TCP	0.0.0.0:5357	X270:0	LISTENING
TCP	0.0.0.0:17500	X270:0	LISTENING
TCP	0.0.0.0:49664	X270:0	LISTENING
TCP	0.0.0.0:49665	X270:0	LISTENING
TCP	0.0.0.0:49666	X270:0	LISTENING
TCP	0.0.0.0:49667	X270:0	LISTENING
TCP	0.0.0.0:49668	X270:0	LISTENING
TCP	0.0.0.0:49669	X270:0	LISTENING
TCP	127.0.0.1:943	X270:0	LISTENING
TCP	127.0.0.1:5939	X270:0	LISTENING
TCP	127.0.0.1:17600	X270:0	LISTENING
TCP	127.0.0.1:50525	view-localhost:50526	ESTABLISHED
TCP	127.0.0.1:50526	view-localhost:50525	ESTABLISHED
TCP	127.0.0.1:63638	view-localhost:5939	ESTABLISHED
TCP	127.0.0.1:63655	view-localhost:63656	ESTABLISHED
TCP	127.0.0.1:63656	view-localhost:63655	ESTABLISHED
TCP	127.0.0.1:63692	view-localhost:63693	ESTABLISHED
TCP	127.0.0.1:63693	view-localhost:63692	ESTABLISHED
TCP	127.0.0.1:63698	view-localhost:63699	ESTABLISHED
TCP	127.0.0.1:63699	view-localhost:63698	ESTABLISHED
TCP	127.0.0.1:63703	view-localhost:63704	ESTABLISHED
TCP	127.0.0.1:63704	view-localhost:63703	ESTABLISHED
TCP	192.168.0.99:139	X270:0	LISTENING
TCP	192.168.0.99:64009	162.125.34.6:https	CLOSE_WAIT
TCP	192.168.0.99:64128	204.79.197.222:https	ESTABLISHED
TCP	192.168.0.99:64137	13.107.42.254:https	ESTABLISHED
TCP	192.168.0.99:64138	131.253.33.254:https	ESTABLISHED
TCP	192.168.0.99:64139	204.79.197.254:https	ESTABLISHED
TCP	192.168.0.99:64142	148.61.30.132:https	ESTABLISHED
TCP	192.168.0.99:64143	148.61.30.132:https	ESTABLISHED
TCP	[::]:135	X270:0	LISTENING
TCP	[::]:445	X270:0	LISTENING
TCP	[::]:5357	X270:0	LISTENING
TCP	[::]:17500	X270:0	LISTENING

Figure 12: Netstat with uplink disabled

6. Conclusions

In our scenario we were able to provide some evidence of a method that could be used in conjunction with other rules and policies to detect rogue switches connected to a network. We were able to determine based on the port link state that there was another device, most likely a rogue switch between the good switch and the computer. The fact that the computers ethernet connection never dropped even though we severed the link from the good switch to a port suspected of having multiple MACs shows that there was a device on the other end keeping the ethernet connectivity of the computer alive. Pulling all the information together from the Windows Logs, Wireshark, and Windows command line tools we formed a good method of what to look for on a computer to see if there is a rogue switch installed somewhere on the network that may affect network connections and protocols.

Another method that could be used that is related to our findings is using the DHCP server, if one exists. You could use DHCP server logs to see if that machine starts asking for a new lease after the good network port is toggled on and off, essentially turning off the uplink on the rogue switch. If the computer doesn't ask for a new

lease, it's likely that the computer didn't see the port state change, which means something blocked it from seeing that, most likely a rogue switch.

If an adversary was able to install a switch that was set to drop the port link states on its own ports when its uplink port also dropped off, that could prove to be another problem in of itself. Overall, we believe the findings of this scenario show another novel method that could be used to detect rogue switches on a network.

References

- Bhuse, Vijay and Kalafut, Andrew and Dohn, Lisa. (2019) "Detection of a Rogue Switch in a Local Area Network", The Fourteenth International Conference on Internet Monitoring and Protection. ICIMP 2019. ISBN: 978-1-61208-729-0.
- James, Vineet, "Network Automation Methodology for Detecting Rogue Switch" (2019). Technical Library. 337. <https://scholarworks.gvsu.edu/cistechlib/337>
- Keohane, S. M. McBrearty, G. Mullen, S. Murillo, J. and Shieh, J. (2011, August 2). US7991877B2 - Rogue router hunter. Retrieved March 1, 2019, from <https://patents.google.com/patent/US7991877B2/en>
- Lin, C.L. and Chen, T.H. (2011, December 6). US8074279B1 - Detecting rogue access points in a computer network. Retrieved March 1, 2019, from <https://patents.google.com/patent/US8074279B1/en>
- Netgear Networking Products. (2021). Retrieved 11 October 2021, from <https://www.netgear.com/>
- Services, P. (2021). Managed Versus Unmanaged Switches. Retrieved 11 October 2021, from <https://www.cisco.com/c/en/us/products/switches/what-is-a-managed-switch.html>
- Solarwinds. Detecting and Preventing Rogue Devices. Retrieved March 1, 2019, from <https://www.solarwinds.com>
- Wireshark. Go Deep. (2021). Retrieved 11 October 2021, from <https://www.wireshark.org/>

Aligning South African Data and Cloud Policy with the PoPI Act

Emma Raaff, Nicole Rothwell and Aidan Wynne

CAIR, School for Data Science and Computational Thinking, Stellenbosch University, South Africa

elraaff@gmail.com

ngrothwell00@gmail.com

aidanwynne.27@gmail.com

Abstract: On 1 April 2021, the South African government released the Draft National Data and Cloud Policy for public comment. This policy aims to support the digital economy in South Africa by implementing initiatives to augment the development of digital infrastructure and skills, with a specific focus on cloud computing. The adoption of the policy would help position the country as a data-driven economy. However, the implementation of such a policy is predicated on the existence of supporting digital infrastructure and aligned inter-departmental goals. Currently, there are many technological and legal considerations that must be addressed in order for such a policy to be implemented successfully. One particularly important consideration is how this policy relates to South Africa's Protection of Personal Information Act, which came into effect on 1 July 2021. The Act sets out legal imperatives for the collection, storage and use of personal information belonging to South African citizens. The aim of this paper is to analyze how the Draft National Data and Cloud Policy relates to the Protection of Personal Information Act in terms of what is proposed in the Policy and what is legislatively imperative in the Act. The overarching context for this evaluation is data governance in the cloud with an emphasis on the security of personal information. Macroeconomic threats and a shortage of critical ICT skills presupposes technical challenges for the implementation of a cloud service. Formal Concept Analysis is utilized to conceptualize and understand the relationships between the Draft National Data and Cloud Policy and the Protection of Personal Information Act. Classification, Confidentiality and Open Data are presented as technological challenges to cloud implementation. This paper aims to contribute towards an understanding of these challenges and how they are affected by South Africa's legislation and policies. We therefore investigate how the Draft National Data and Cloud Policy is situated in the broader context of data protection in South Africa.

Keywords: Cloud data governance, Cloud computing, Digital security, The PoPI Act, Data and Cloud Policy

1. Introduction

The current landscape of cloud computing in South Africa necessitates appropriate governance strategies to materialise the benefits associated with the Fourth Industrial Revolution. The Draft National Data and Cloud Policy (Draft Policy) presents an opportunity to investigate how the establishment of an open digital economy will align with existing policies, legislation and regulations. This research paper aims to analyse the current data protection landscape in South Africa to determine whether it is able to accommodate what is proposed by the Draft Policy. Insights drawn from a Formal Concept Analysis (FCA) lattice will be utilized to evaluate the relationship between the Draft Policy and the PoPI Act. This paper aims to offer recommendations for the restructuring of the Draft Policy to aid in its implementation.

The next section provides some background to contextualise the paper. Section 3 outlines the research design, after which Section 4 displays the results of the FCA process. Thereafter, Section 5 provides critical evaluation and discussion of results.

2. Background

Data governance encompasses the rights and responsibilities associated with the management of data assets (Al-Ruithe, Benkhalifa & Hameed, 2018). Forward-thinking organizations consider the implementation of effective data governance as a means to solve data-related issues (Al-Ruithe, Benkhalifa & Hameed, 2018). Cloud governance involves policies relating to the security, privacy, availability, compliance and location of cloud services. Since data governance is under-researched and often applied informally, regulations surrounding the management of data require enhanced clarity if they are to be effectively implemented.

The association between data and cloud computing is becoming increasingly prominent with the global drive towards digitisation. To protect individuals, organizations and governments, the creation of functional data and cloud governance policies and frameworks is imperative. However, there is limited research on cloud governance. A well-constructed governance strategy provides vision, establishes accountability, and encourages desirable behaviour from role-players in the data and cloud domain (Al Ruithe, Benkhalifa & Hameed, 2018).

2.1 Landscape on Cloud Computing South Africa

According to Kumar & Kumar (2016), “cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.” It depends on a well-established ICT ecosystem, contingent on broadband infrastructure, to ensure the availability and reliability benefits of cloud computing.

Cloud computing has the potential to reduce expenditure on physical ICT infrastructure and can augment government initiatives to provide more affordable service delivery (Gillwald & Moyo, 2017). Since 1994, the South African government has made significant progress in addressing ICT issues in the country. Effort has been made to achieve greater coordination between government departments and to align ICT with business processes (Gillwald & Moyo, 2017:44). However, challenges to the implementation of an effective data and cloud policy are still present.

2.2 Draft Policy Overview

Digitization presents South Africa with an opportunity to address its socio-economic challenges (Draft National Policy on Data and Cloud, 2021). Government and private sector organizations are increasingly shifting their service delivery models to the digital domain to provide accessible services to citizens and customers. The creation of the Draft Policy has been motivated by the acknowledgement that South Africa’s success in overcoming economic challenges is largely dependent on exploiting opportunities in the digital economy.

Currently, data generated by government is stored mostly in privately-owned databases. The lack of policy to guide data acquisition, ownership, storage, use and analytics pose a threat to national security and economic growth. There are numerous legal, policy, and regulatory documents situated in separate government departments, resulting in a lack of coordination. The Draft Policy aims to align the disparate documents in such a way that enables the country to develop a competitive digital economy.

The Draft Policy stipulates that the State Information Technology Agency (SITA) should act as the cloud champion. The Policy also emphasizes skills and infrastructure development. Therefore, the Policy targets some of the crucial developmental issues that South Africa faces.

2.3 The Landscape of Data Protection

New threats to the integrity and confidentiality of personal information are introduced by the Internet of Things (IoT), cloud computing and Big Data, and data protection initiatives are required as a result (Botha et al, 2017). Globally, over one hundred countries have implemented data protection laws. The General Data Protection Regulation (GDPR) is arguably considered the gold standard of privacy legislation. South Africa’s PoPI Act has been formed based on GDPR principles, alongside models from the USA, Canada, Australia and the UK (Botha et al, 2017).

2.4 The PoPI Act Regulation

The PoPI Act provides directives for the storage and processing of personal information belonging to South African citizens and juristic entities. Every responsible party that collects, stores, modifies or uses information must comply with the PoPI Act and will be held accountable for non-compliance. The legislation aims to protect Personal Information (PI), prevent theft, and uphold the fundamental human right to privacy (Nel, 2021). For the Act to effectively meet its aims, there are legal consequences for non-compliance, including civil class action, fines of up to R10 million and jail time of up to 10 years (Botha et al., 2017).

The PoPI Act outlines key stakeholders, defined in Table 1 (Republic of South Africa, 2013).

Table 1: Key Role Players in the PoPI Act

Key Role Players in the PoPI Act	
Actors	Role definition
Information Regulator	A government institution that investigates and fines responsible parties, enforcing compliance with the Act
Data Subject	Any party to whom personal information belongs
Responsible Party	A public or private body who decides why and how to process information
Data Processor/Operator	A party that processes personal information on behalf of the responsible party

3. Research Design

What the Draft Policy envisions appears to challenge what is prescribed by the PoPI Act and what is practically possible to implement within the South African context. Four research questions were formulated to guide the discussion surrounding data governance in South Africa. The aim of these questions is to understand the most important considerations for data governance and cloud security by making explicit the relationships between the PoPI Act and the Draft Policy. These questions contribute towards the overarching aim of understanding the current data protection landscape and evaluating whether South Africa is in a position to accommodate what is proposed in the Draft Policy.

3.1 Research Questions

The questions guiding the research problem include the following:

Q1: Does the Draft Policy have implications for the confidentiality of data?

Q2: What does the Draft Policy stipulate regarding the regulation of data and how does this relate to the PoPI Act?

Q3: Are there contradictions in what the Draft Policy proposes and how does the PoPI Act impact these discrepancies?

Q4: How does the Draft Policy propose the security of digital information assets and adhere to the PoPI Act?

3.2 Methodology

The research method utilised is Formal Concept Analysis (FCA), which is primarily used for graphical knowledge representation and exploration (Priss, 2007). FCA involves the representation of objects and attributes in a formal concept lattice, which can be analysed to discover concepts and relationships from the data. This paper utilises FCA to understand relationships between the Draft Policy and the PoPI Act.

Objects, attributes and the relationships between them are captured in a cross table, known as the formal context. The cross table is then used to build a formal concept lattice. In the lattice, each node represents a formal concept, consisting of a single object or several closely related objects and their shared attributes. Connections between concepts are represented by lines. For the purposes of this paper, aspects of the Draft Policy were encoded as objects and the principles from the PoPI Act were encoded as attributes. The lattice visually depicts the relationships between objects from the Draft Policy and attributes from the PoPI Act.

3.3 Application of FCA to the Draft Policy and the PoPI Act

3.3.1 Data Processing

The first step involved extracting the most significant concepts and principles from the Draft Policy and the PoPI Act. This was achieved by encoding policy interventions in the Draft Policy as objects, and key principles in the PoPI Act as attributes. A cross table of objects and attributes was created which formed the formal context.

3.3.2 Data Mapping

The formal context was used to build a formal concept lattice, which was created automatically by the selected FCA tool, ConExp. The lattice visually depicts the relationships between objects and attributes by systematically arranging them in such a way that clearly displays the paths between concepts. Figure 1 depicts the concept lattice that was created using the formal context.

a Data Regulator would raise implications for the continued existence of other regulatory authorities. The PoPI Act has its own Information Regulator, and therefore a significant issue in the Draft Policy's mandate relates to how the PoPI Act will be upheld with the new role of a Data Regulator.

The lattice provides insight into the importance of the Data Regulator in the Draft Policy. The Data Regulator object is positioned close to the top of the lattice, subsuming many objects below it. Therefore, the Data Regulator is connected to, and has implications for, many interventions in the Draft Policy. However, despite being of high importance within the Policy, there is little discussion regarding the details of the Data Regulator's appointment and mandate in terms of the PoPI Act, especially regarding implications for the Information Regulator. Monitor and Enforce Compliance is the only attribute attached to the formal concept associated with the object of the Data Regulator in the lattice. Monitor and Enforce Compliance is synonymous with the role of the Information Regulator. An observation from the lattice is therefore the conservative association of the Data Regulator's responsibilities with the PoPI Act. This illustrates the importance of clarifying exactly how the Policy, and particularly the Data Regulator, relates to the PoPI Act's Information Regulator. The Draft Policy's proposition for a Data Regulator is to consolidate the regulatory bodies stipulated in relevant legislation, suggesting that the Data Regulator could replace these bodies completely. However, there is little information provided on how exactly the Data Regulator would integrate with the PoPI Act and whether this new body would, in fact, entirely replace the Information Regulator.

It is also unclear whether there is capacity to establish such a body when considering that the Information Regulator under the PoPI Act faced significant personnel shortfalls; even after its implementation in July 2021, the body did not have a Chief Executive Officer (Nel, 2021). There is also uncertainty surrounding the implementation of the PoPI Act, with organisations being unsure how to comply with the Act due to the fact that the responsibility for educating organisations on compliance falls to the Information Regulator. The challenges experienced by the Information Regulator are a concern when considering the establishment of the Data Regulator, as this body would require more resources and capability than the Information Regulator due to its responsibility for overseeing all aspects of data governance.

4.3 Question 3: Potential Contradictions

The primary rationale behind an open data economy model and an open data strategy is to provide an inclusive environment that promotes innovation and economic growth (Draft National Policy on Data and Cloud, 2021). The Draft Policy argues for data localization and residency, in order to ensure that the country receives value from local data processing. There are potential contradictions between the Draft Policy's stipulations regarding its open data policy and those regarding data localization and residency, which could infringe on the rights of citizens.

The concept lattice supports the idea of potential contradictions between the Draft Policy and the requirements of the PoPI Act. The attributes of International Privacy Protection Guidelines, Information Sharing, and Cross-Border Cooperation, from the PoPI Act, are all related to the three most significant objects, of which Open Data is the most relevant when evaluating the Open Data Strategy proposed by the Draft Policy. Open Data is an object of the most general formal concept in the lattice, having no attributes and subsuming all other concepts.

Data localization, which is the most restrictive policy measure referenced in the Draft Policy, stipulates that all data created within a nation are the property of that nation (Wong, 2020). This results in the nation having ownership over the data, whereby its governing laws dictate how the data will be treated and whether cross-border data flow is allowed. The PoPI Act's primary purpose is set out in Section 2. S 2(a)(ii) recognizes that privacy is a fundamental right, and that safeguards to protect this right must be in place.

Furthermore, S 2(b) of the Act aims to regulate how personal information is processed by establishing conditions that are in alignment with international standards and best practices. The purpose statements set out in S 2 of the PoPI Act support the Draft Policy's open data strategy and the free flow of information in South Africa and across its borders. The mandate also maintains regulatory consonance with international processing standards regarding personal information (Republic of South Africa, 2013). However, one can argue that the South African government could introduce a data localization framework that contradicts the open data strategy proposed in the Draft Policy. Restricting the free flow of data will create a closed data environment. Therefore, while aspects of the PoPI Act and Draft Policy support an open data strategy, the proposal of a data residency, sovereignty,

and localization framework will conflict with the open data strategy as stipulated in the Draft Policy. Ambiguity surrounding this issue should be removed by providing clarity in the Draft Policy.

4.4 Question 4: Security Issues

As the nation undertakes digitalization through the development of state digital infrastructure, security threats are presented. Cybersecurity is an element that requires consideration by the South African Government due to the implications of the Draft Policy (Sutherland, 2017). South Africa has fallen behind in prioritizing cybersecurity legislation and lacks the skilled labor to meet the demands required of the nation. South Africa's cybersecurity landscape, coupled with the risks associated with cloud computing, raises the question of whether South Africa has the ability to sufficiently mitigate cyber threats.

The concept lattice indicates all objects and attributes related to the Digital Security Enforcement object. The lattice can be used to substantiate cybersecurity shortfalls that the Draft Policy and the PoPI Act present. This formal concept, Digital Security Enforcement, is a very specific concept which is subsumed by most of the other concepts in the lattice. The main attribute intents of Digital Security Enforcement that will be evaluated are Report Security Compromises, Monitor and Enforce Compliance, and International Privacy Protection Guidelines.

The Draft Policy identifies two important considerations under S 10.3, recognizing the role of the PoPI Act and its guiding principles while ensuring accurate monitoring and enforcement of compliance (Draft National Policy on Data and Cloud, 2021). The second issue, under S 10.5 recognizes that South Africa has among the worst "online crime" in the world (Draft National Policy on Data and Cloud, 2021). One of the main concerns of the Draft Policy's open data strategy is that the proposed data protection and cybersecurity policy interventions may not suffice in safeguarding the proposed digital infrastructure. The PoPI Act also carries implications for cybersecurity, given its broad exemptions for national security (Sutherland, 2017). Government and private entities have not expressed confidence in the current cybersecurity infrastructure to prevent cyber-attacks. This raises doubt regarding whether the Draft Policy and its associated open data strategy can be implemented without exacerbating vulnerabilities.

5. Evaluation and Discussion

5.1 Question 1: Confidentiality

S 19 of the PoPI Act states that the integrity and confidentiality of personal information must be upheld while mitigating unauthorized access and unlawful processing. The Draft Policy proposes the instatement of a public High-Performance Computing and Data Processing Centre (HPCDPC). This will make data widely accessible, which fulfills the Policy's open data strategy. This public cloud will store all non-sensitive government and state data. Existing data protection laws and minimum-security standards will determine what is classified as non-sensitive data. Sensitive data will be stored in a private cloud to ensure adequate measures are in place to ensure confidentiality. Ensuring the confidentiality of data first requires accurate classification frameworks. The Draft Policy states that any handling of personal information will be in accordance with the PoPI Act. However, this rests on classification processes. Furthermore, if the Draft Policy were to be implemented, it would require classification activities to be conducted with collective responsibility and a high regard for best practice. There are concerns regarding the ability of current classification frameworks to uphold the PoPI Act if the Draft Policy were to be implemented. Before the Draft Policy is implemented, existing cybersecurity legislation, frameworks and initiatives surrounding classification should be evaluated to ensure that they offer adequate protection of confidential personal information.

5.2 Question 2: Regulation

A prominent point of ambiguity in the Draft Policy is found in s 10(10.6.2), which proposes a "review of existing regulatory authorities" and the subsequent establishment of a Data Regulator. It is unclear whether the Data Regulator would completely replace existing regulatory bodies stipulated in the PoPI Act and other legislation, or if it would extend and support existing bodies. There is also significant concern around the Data Regulator receiving sufficient funding for the responsibility placed on it, as it would oversee all aspects of data governance and management. This concern arises primarily out of difficulties faced by similar bodies such as the PoPI Act's Information Regulator.

A solution to these challenges is for the Draft Policy to provide a more explicit and detailed description of what exactly the Data Regulator would do and how it would integrate with existing security and data-related regulatory bodies. In addition, insight is needed into how the review of existing regulatory bodies would be conducted and therefore how the Data Regulator would come into existence. This would prevent ambiguity and uncertainty. The Draft Policy must also ensure that research is undertaken into the difficulties faced by similar regulatory bodies in order to reduce the challenges presented to the Data Regulator.

5.3 Question 3: Potential Contradictions

The Draft Policy states that data ownership and cross-border data flow will be in accordance with the PoPI Act (Draft National Policy on Data and Cloud, 2021). While this reduces the risk of ownership and data flow being in contradiction to the PoPI Act, there are issues to consider regarding the way that the PoPI Act and the Draft Policy deals with ownership and data flow. Malinga (2021) reports on these issues and raises the important point that static or raw data lacks economic value. Value is generated through processing and data flows. While the Draft Policy recognises the need to ensure that data increases in value through effective processing and transfer, there appears to be significant emphasis on keeping data processing local under the PoPI Act. Therefore, while there is no outright contradiction between the Draft Policy's aims and the PoPI Act's provisions regarding data ownership and flow, there is a risk that this approach to data ownership and cross-border flow, drawn from the PoPI Act, contradicts the Draft Policy's own aims to develop the digital economy to its full potential. The Draft Policy should then be clearer in how the capacity of local data processing will be improved to ensure that it can fulfil its aims of creating increased value to develop a digital economy.

With regards to non-personal information, it could prove effective to allow freer flow of data across borders and enable non-local data value generation instead of creating a situation of data localisation. As it stands, it seems that there is a risk for this under the Draft Policy. This could result in the loss of potential value that could be generated through non-local data processing and a freer flow of data across borders. Therefore, while the Draft Policy places emphasis on developing local data processing and value creation, it must be ensured that South African public and private institutions have the required faculty to generate this value. It should also be ensured that the opportunity for value creation provided by non-local processing and ownership of data is not overlooked.

A recommendation for solving the abovementioned issues is to balance local data processing and value creation with non-local data processing, flow and ownership. The Draft Policy should also ensure that sufficient research is conducted into the potential for South African institutions to process data and produce value, and use this research to inform how restrictions placed on cross-border data flow and ownership are formulated.

5.4 Question 4: Security Issues

The Draft Policy indicates a need for Cybersecurity Measure reform and improvement as the current cybersecurity ecosystem in South Africa is arguably unequipped to mitigate new threats. The protection of personal information from cybercrime is dependent on South Africa's capability to implement an appropriate risk mitigation infrastructure. This has been a priority for the country in recent years; Gillwald & Moyo (2017) note that SITA has been working to improve its security standards in efforts to mitigate cybercrime. However, due to limited capacity, it is unclear whether the state of cybersecurity in South Africa is currently prepared to mitigate cyber threats at the level required by the Draft Policy. The Cyber-security Bill is arguably considered as a turning point for South Africa's cybersecurity initiatives. However, it is not referred to explicitly in the Draft Policy. A recommendation is for the Cybersecurity Bill to be acknowledged while stating that the Draft Policy will be in accordance with the Bill to emphasize how cybersecurity legislation is being enforced, thereby ensuring that there is legislated protection against unlawful activity.

The South African government should first attempt to improve the state of cybersecurity in the nation before attempting to implement the Draft Policy. The Draft Policy places emphasis on this, stating that current cybersecurity measures should be updated. For example, s 10(10.5.2) and s 10(10.5.3) state that cybersecurity legislation such as the NCPF should be reviewed to ensure that it accounts for the threats and risks accompanying increased digitization, as well as ensuring capability to respond to risks. However, cybercrime is presented as a significant challenge and it is unclear whether there is sufficient capacity to develop the current cybersecurity infrastructure to a point where the Draft Policy can be implemented without exacerbating vulnerabilities. S 10(19.1.5) states that the HPCDPC will be cloned to establish two similar centres which will provide backup and ensure operation continuity in the case where the HPCDPC is a target of a cyberattack. The Draft Policy's

contingency plan accounts for the current cybersecurity climate, however, legislation and frameworks should be reviewed as a precursor to the infrastructural development under the Policy to first create a secure cybersecurity ecosystem. A recommendation is that more should be invested into developing the country's cybersecurity infrastructure and skills before the Draft Policy is implemented, in order to ensure that South Africa is fully equipped to deal with the new threats and risks associated with cloud computing and increased digitization.

5.5 Establishing SITA as the Cloud Champion

The Draft Policy stipulates that SITA will become the champion of cloud computing in the country. A significant emphasis in the Draft Policy is the establishment of the High-Performance Computing and Data Processing Centre (HPCDPC) which will provide data processing facilities and cloud computing services. The Draft Policy places much of the responsibility for this Centre on SITA. However, implementing the HPCDPC will prove challenging, since SITA faces significant obstacles that could hinder its ability to successfully carry out its precept under the Draft Policy (Gillwald & Moyo, 2017). SITA's 2020-2025 Strategic Plan (n.d.:42) highlights the challenges still faced by the agency. Some of the key issues highlighted in the Strategic Plan relate to funding, leadership and corruption. Therefore, SITA will face challenges meeting the requirements associated with a cloud champion. In the past, SITA has prohibited the use of cloud computing services, which has resulted in the constriction of this technology at a national and provincial government level (Gillwald & Moyo, 2017). Thus, a fundamental shift is required by SITA if they are to become the primary custodian of cloud computing under the Draft Policy. A solution to ensure that the agency carries out their role effectively is to ensure accountability to an external body. If responsible parties, and organizations in particular, are going to entrust personal information for which they are responsible to SITA, there must first be stringent accountability and security practices in place. Therefore, SITA is a crucial component when evaluating the likelihood for success of the Draft Policy.

6. Conclusion

The Draft National Data and Cloud Policy is a promising development in South Africa's digital environment. While the policy is recognized as being valuable, there are many important considerations that have to be accounted for regarding its implementation. This paper has focused on several areas in which the Draft Policy relates to the PoPI Act, including availability and confidentiality, data regulation, data localisation and digital security. In addition, a discussion of SITA as the cloud champion was presented. An evaluation of potential challenges in those areas was conducted in an attempt to contribute to an understanding of data governance in the cloud. Suggested strategies to address the implications of the Draft Policy were presented. The evaluation presented in this paper was guided by Formal Concept Analysis and involved the examination of a concept lattice generated by ConExp. The Draft Policy promotes a modern data governance landscape in South Africa. However, more transparency and clarity in the policy is required to ensure that the principles of the PoPI Act are effectively complied with, along with the restructuring of the Draft Policy document to enable the benefits associated with cloud computing to be realized.

Future work could include the evaluation of relationships between the Draft Policy and other legislation such as the National Cybersecurity Policy Framework. As the Draft Policy is still in draft format, once implemented, future work could focus on the impact of the Policy in comparison to how other countries have undertaken the adoption of cloud computing.

Acknowledgements

We would like to sincerely thank Arina Britz and Bruce Watson for their guidance and support, while encouraging us throughout. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number:136211) and the Centre for AI Research (CAIR).



References

Ahuja, S. and Mani, S., 2012. Availability of Services in the Era of Cloud Computing. *Network and Communication Technologies*, 1(1).

- Al-Ruithe, M., Benkhalifa. and Hameed, K. (2018). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing* [Electronic], 23(5). Available: <https://doi.org/10.1007/s00779-017-1104-3> [2021, October 5].
- Botha, J., Grobler, M.M., Hahn., J. and Eloff, M.M. (2017). A High-Level Comparison between the South African Protection of Personal Information Act and International Data Protection Laws. *International Conference on Cyber Warfare and Security (ICCWS)*.
- Department of Communications and Digital Technologies. (2021). *Draft National Policy on Data and Cloud* [Online]. Available: https://www.gov.za/sites/default/files/gcis_document/202104/44389gon206.pdf [2021, October 5].
- Gillwald, A and Moyo, M. (2017). *Modernising the Public Sector through the Cloud*. [South Africa]: Africa Portal [Online]. Available: https://media.africaportal.org/documents/Cloud-Computing-in-the-public-sector-final-25052017_V03.pdf [2021, October 5].
- Kandeh, A., Botha, R. and Fitcher, L. (2018). Enforcement of the Protection of Personal Information (POPI) Act: Perspective of data management professionals. *SA Journal of Information Management*, 20(1).
- Kumar, M and Kumar, N. (2016). Cloud Computing: Architect and Use of Technology in the Past, Present and Future. *International Journal of Research in Engineering, Technology and Science* [Electronic], 6. Available: <http://www.ijrets.com/wp-content/uploads/2016/08/160314084-11.pdf> [2021, October 5].
- Malinga, S. (2021). *Several flaws in SA's draft data and cloud policy, say experts* [Online]. Available: <https://www.itweb.co.za/content/xA9PO7NZ46Z7o4J8> [2021, October 5].
- Nel, P. (2021). Personal Interview. 22 July.
- Priss, U. (2007). Formal Concept Analysis in Information Science. *Annual Review of Information Science and Technology* [Electronic], 40(1). Available: <https://doi.org/10.1002/aris.1440400120> [2021, October 5].
- Republic of South Africa. (2013). *Protection of Personal Information Act*. No 4.
- State Information Technology Agency. (n.d.). *2020-2025 Strategic Plan*. [Pretoria]: SITA [Online]. Available: <https://www.sita.co.za/sites/default/files/Strategic%20Plan%202020-2025.pdf> [2021, October 6].
- Sutherland, E. (2017). Governance of Cybersecurity – The Case of South Africa. *The African Journal of Information and Communication*, (20): 83-112.
- Thamm, M. (2018). SAPS/SITA capture: FDA holds South Africa to ransom, threatens to 'collapse criminal justice system'. *Daily Maverick* [Electronic], 5 April. Available: <https://www.dailymaverick.co.za/article/2018-04-05-sapssita-capture-fda-holds-south-africa-to-ransom-threatens-to-collapse-criminal-justice-system/> [2021, October 6].
- Wolff, K.E. (1993). A First Course in Formal Concept Analysis How to Understand Line Diagrams. *SoftStat '93: Advances in Statistical Software 4*, 429-438 [Electronic]. Available: <https://sites.tufts.edu/ancientbirds/files/2018/06/a-first-course-in-formal-concept-analysis.pdf> [2021, October 5].
- Wong, B. (2020). Data Localization and ASEAN Economic Community. *Asian Journal of International Law* [Electronic], 10(1). Available: <http://dx.doi.org.ez.sun.ac.za/10.1017/S2044251319000250> [2021, October 5].

Transient Execution and Side Channel Analysis: a Vulnerability or a Science Experiment?

Michael Shepherd, Scott Brookes and Robert Denz

Riverside Research, Lexington MA, USA

shepherdm1@wit.edu

sbrookes@riversideresearch.org

rdenz@riversideresearch.org

Abstract: In the world of computer security, attackers are constantly looking for new exploits to gain data from or control over a computer system. One category of exploit that can prove quite effective at accessing privileged data is side channel exploits. These exploits attempt to take advantage of vulnerabilities that are inherent in the design of a system rather than vulnerabilities in the code that has been written for and is running on said system. In other words, they exploit side effects of computation. Examples of this include measuring the power consumption of a system's processor over time and analysing that power usage to leak system secrets or reading secrets from a system by analysing the electromagnetic radiation the system leaks as it processes data. Another type of side channel attack is a cache-based side channel attack, which exploits the timings of cache and memory accesses to determine data from the target system. We discuss some of the more common types of side channel attacks used to interpret data values from the microarchitectural changes created by transient executions. In particular, we will focus on attacks that are capable of recovering data that is processed through transient execution in some way and then wrongly accessed using a side channel, such as the Spectre and Meltdown classes of attack. We also discuss other attacks of a similar type and survey some popular mitigations for these attacks. We provide a survey of all available Spectre proof-of-concept repositories on GitHub, evaluating whether they work on different platforms. Finally, we review our experiences with these types of attacks on modern systems and comment on the attacks' practicality, reliability, and portability. We conclude that these types of attacks are interesting, but there are some practicality and reliability concerns that make other attacks easier much of the time.

Keywords: Side Channel, Transient Execution, Spectre, Meltdown, Prime and Probe, Flush and Reload

1. Introduction

Side channel attacks utilize by-products of computation such as power, electromagnetic (EM) radiation, or timing to leak information about the computation. Most recently, the use of timing side channels alongside transient execution has led to new side channel vulnerabilities.

The transient domain refers to a type of instruction executed by the processor before it is known whether the instruction should be executed. One example is speculative execution, where the processor might use its Branch Target Buffer (BTB) to guess which direction a branch will take and transiently execute instructions on that path. Instructions executed in this transient domain do not make changes to the visible processor state unless they are committed. If the processor determines that the transient instructions are not valid, it discards the results of the instructions.

Although transient execution does not change architectural processor state until the execution is validated, it does leave traces in the microarchitectural state. For example, transient instructions often change the state of the cache by evicting or caching certain data. Cache-based side channels can then recover this microarchitectural state. As a result, program secrets that are operated on in the transient domain can be recovered from the microarchitectural state of the processor by attackers using a side channel.

In this paper, we examine different types of attacks that expose secrets in the transient domain and recover those secrets using cache-based timing side channels. As this class of attacks does not require physical access to the machine, they are especially potent.

The main contributions of this paper include:

- Comparing and contrasting different cache-based timing side channel attack methodologies.
- Explaining how various transient execution attacks work.
- Providing a thorough analysis of the Spectre proofs-of-concept available in the public domain.
- Demonstrating a novel attack, SpectreXP, which uses a network connection to exploit Spectre across the process isolation boundary.

In Section 2, we provide a brief discussion of background and related work. In Section 3, we explain cache-based timing side channels and study methods for recovering secrets from the cache using timing information. Then, in Section 4, we examine different types of attacks that reveal secrets in the transient domain. After these explanations, we will discuss our practical experience during our experimentation with these types of attacks on different types of real systems in Section 5. Section 6 offers concluding thoughts and future work.

2. Background & Related Work

Transient execution attacks turned the computer security world upside down in 2018 with the publication of the original Spectre (Kocher et al, 2019) and Meltdown (Lipp et al, 2018) attacks. The gravity of the situation became clear when it was shown that Spectre could be exploited via Javascript to break out of a browser sandbox (Novack and Reichert, 2018).

Patches for the original vulnerabilities came out relatively quickly. Browsers reduced the precision of their timers to interrupt the cache-based timing side channels used to recover secrets in the Spectre attack (Wagner, 2018) while operating system vendors implemented patches such as KPTI that mitigated Meltdown (kernel.org; Gruss, 2017).

Unfortunately, patching this class of vulnerability turned out not to be so easy. Not only was it shown that reducing timer precision was not sufficient to stop Spectre (McIlroy et al, 2019) but a variety of additional instances of this class of attack began to surface including Foreshadow (Van Bulck et al, 2018; Weisse et al, 2018), ZombieLoad (Schwarz et al, 2019), RIDL (Van Schaik et al, 2019), Fallout (Canella et al, 2019), and others. Indeed, Canella et al. (2019) provide an analysis showing that there are tens of instances of this class of attack between the Spectre and Meltdown approaches alone.

This leads us to the current state of the art. Defenders are playing a game of whack-a-mole as researchers are still bringing new attacks that exploit the leakage from the microarchitecture to the architectural level of the processor. Works that survey this landscape in greater detail than allowed in the scope of this paper include those by Xiong and Szefer (2020), Canella et al. (2019) and Canella et al. (2020). While these works deeply evaluate the technical aspects and theory of transient execution attacks, this paper will focus more heavily on our experience with researching currently available exploit examples and testing them on real world hardware and software. We will also discuss our analysis of the limitations and practicality issues that exist that affect the usefulness and danger of transient execution attacks in the wild.

3. Cache-Based Timing Side Channel Methods: Extracting Secrets from the Cache

Two side channel methods are used in a majority of cases to retrieve data from a cache altered by transient execution. These methods are known as prime and probe and flush and reload.

3.1 Prime and Probe

A popular side channel for leaking data from the cache is known as “prime and probe” (Osvik et al, 2006; Tromer et al, 2010) which uses a 3-step process, illustrated in Figure 1, to pull data from the cache without assumptions about shared memory.

1. First, the attacker “primes” the cache, filling it with random attacker-controlled data. This ensures that if the attacker attempts to access any of this data, it will result in a cache “hit”, or a very fast access of the requested data as it is retrieved from cache instead of the significantly slower RAM.
2. The attacker then waits for the victim to execute code that accesses the target data. This process of pulling data into the cache causes some of the attacker-controlled data to be evicted from the cache.
3. Finally, the attacker “probes” the cache by attempting to access its attacker-controlled data, looking for a cache “miss” where data was evicted in step 2. In other words, it looks for a slow data access taking longer to access than it would to retrieve that data from the cache memory.

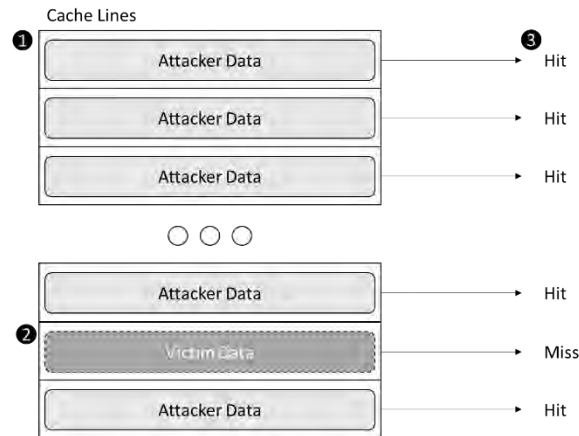


Figure 1: Prime and Probe uses 3 steps: 1 - the attacker fills the cache with known data, 2 - the victim accesses the secret, evicting some attacker-controlled data, 3 - the attacker looks for a miss to determine which data was accessed by the victim

By mapping certain cache lines to certain values, the value of the accessed data can be determined based on which cache line was evicted in the second step. This allows for the extraction of data, including data that is accessed only transiently, from a system without the attacker directly accessing that data using only changes in the cache. This can even be deployed against the Last Level Cache (LLC) as shown by Liu et al. (2015).

3.2 Flush and Reload

A newer type of cache timing side channel called “flush and reload” (Yarom and Falkner, 2014) seeks to improve upon the speed and accuracy of prime and probe with the trade-off of requiring that the victim and attacker processes share memory. One way to achieve this shared memory would be running the attacker and victim process hyperthreaded on the same core, sharing their L1 and L2 cache. However, you can also leverage the last level cache to read data from another process, as it is shared across all cores as shown in Figure 2.

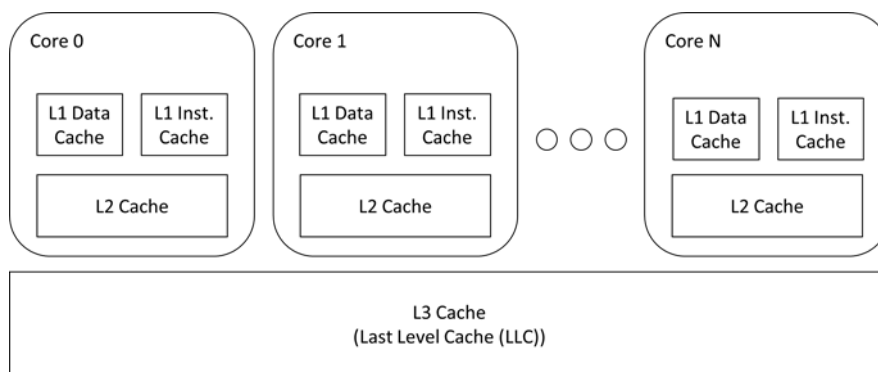


Figure 2: While each core has its own L1 and L2 cache, the L3 cache or “last level cache” (LLC) is shared among all cores, allowing for attacks using flush and reload to work across cores using the LLC as shared memory

Flush and reload is based on the same principles as prime and probe, except it is looking for a cache hit instead of a cache miss, as illustrated in Figure 3.

1. The first step of a flush and reload attack is to simply flush the entire cache.
2. Once the cache has been emptied, flush and reload, like prime and probe, waits for the victim to access target data, again either transiently or correctly. Instead of evicting attacker-controlled data from the cache, this causes some cache lines in the shared memory to “light up” or become filled with data as the system pulls the requested data into the previously empty cache.

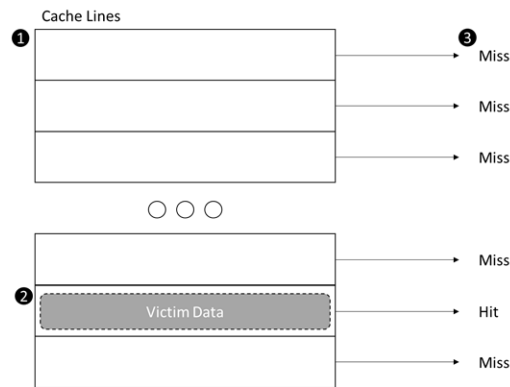


Figure 3: Flush and Reload uses three steps: 1 - the attacker empties the cache, 2 - the victim accesses the secret, filling a cache line with data, 3 - the attacker looks for a cache hit to determine which data was accessed by the victim

1. This attack then “reloads” the cache by going line by line, looking for a cache “hit,” as the lines that have data in them after step 2 will respond to a read much faster than empty lines that must retrieve the requested data from RAM.

4. Transient Execution Attacks: Exposing Secrets in the Transient Domain

While side channel attacks can be used to detect changes made in the cache at the microarchitectural level, that alone is not enough to leak data from a process. Transient attacks make use of transient execution, or execution that is performed ahead of time and thrown out if it is wrong, to leave microarchitectural changes on the cache that can then be picked up by a side channel attack.

4.1 Meltdown

Meltdown (Lipp et al, 2018) leverages a feature known as “out of order execution” in order to overcome memory isolation and read privileged memory, including arbitrary kernel memory. Modern processors use out of order execution to increase performance, as this feature enables the processor to look ahead in the instruction sequence and schedule future operations to idle execution units on the CPU.

Meltdown leverages the fact that while a CPU will throw out data if it was accessed with insufficient permissions, it may transiently load privileged data with an unprivileged program. The CPU throwing out the accessed data ensures normal program execution, but Meltdown leverages the fact that this out of order memory access makes microarchitectural changes to the cache. These changes can be detected and deciphered using a side channel such as flush and reload or prime and probe, bypassing security.

4.1.1 Pseudocode

To illustrate the Meltdown attack, consider the pseudocode shown in Figure 4.

First, we declare a memory region of the same size as the entire cache. Then, we flush the cache. Next, we dereference a kernel address, effectively assigning secret data to the variable `a`. Although this will cause a segmentation fault in the program, it will not crash before the transient domain looks ahead and continues execution to include the operation on the `ath` 64 byte region of `detectArray`. At this point we know that one cache line has been filled with data: the `ath` one corresponding to the secret value. By detecting which cache line results in a hit, one can determine the value of the secret.

Note that the key here is using the secret as an offset within known data. This allows inspection of the known data’s footprint in cache to reveal the value of the secret.

```

char detectArray[256 * 64]; \\ 256 cache lines of 64 bytes each

flush_cache(detectArray);

try {
    char a = *kernel_address;
    detectArray[a * 64]++; \\ run in transient domain
} catch (segfault) {}

for (int i = 0; i < 256; i++)
    if ( is_in_cache(detectArray[i*64]) )
        printf("secret char is %d\\n", i);

```

Figure 4: Code snippet for Meltdown attack from Godbolt (2018)

4.1.2 Meltdown Mitigations

After Meltdown’s discovery, manufacturers and developers rushed to find mitigations to this exploit, as it gives an attacker complete access to a system’s privileged memory. One solution would be to disable out of order execution, but that would result in an unacceptable loss of performance. Other mitigations have high performance overhead as well. One example of this is kernel page table isolation (KPTI) (kernel.org; Gruss, 2017) for the Linux kernel, which creates a shadow copy of the system’s page tables for use when operating in user space mode that do not have kernel space memory mapped in them aside from pointers required for system calls and interrupts. This means that while Meltdown can still leak memory, it cannot leak kernel or physical memory as there are no mappings in the user space page tables for those addresses. KPTI can have a performance impact as high as 30% on certain workloads (Gregg, 2018).

4.2 Spectre

Spectre (Kocher et al, 2019) is another transient execution attack which leverages a feature of modern CPUs called “speculative execution.” Speculative execution improves performance by executing instructions/calculations before they’re needed by the program so they can be ready once they are needed. However, while out of order execution simply looks ahead in the instruction sequence, speculative execution attempts to “guess” what the result of a branch (e.g., an ‘if’ statement) will be and performs operations that may not be required. In this way, if they are required, the execution has already been completed, and if not, the system simply discards the transient execution.

The issue with Spectre is that instructions executed speculatively can bypass security such as bounds checking and wrongly access data within the victim process’s memory address space. To perform a Spectre attack, the attacker trains the CPU to expect a branch to go a particular way by invoking victim code with valid inputs, then invokes said code with a malicious input to make it wrongly speculate and access secret data in the transient domain. Figure 5’s code snippet shows a vulnerable function usable by a Spectre attacker, feeding it valid ‘x’ values to train the branch predictor to expect the if statement to be true, then giving it a malicious x value that results in an out of bounds transient memory access that can be recovered via a side channel attack.

```

if (x < array1_size)
    y = array2[array1[x] * 4096];

```

Figure 5: Spectre gadget

4.2.1 Spectre Mitigations

Spectre is very difficult to mitigate. With Meltdown, simply making the attacker unable to access page tables for sensitive kernel and physical memory means that it cannot leak secrets. The fact that Spectre leaks any data in the victim’s memory address space means that Spectre can leak secrets from within the process itself. Simply disabling speculative execution would defeat Spectre, but this would also result in significant and unacceptable performance loss as modern systems rely heavily on speculative execution for their high performance.

There are some proposed mitigations for Spectre class attacks. One example that Mozilla chose for its Firefox browser was to lower its timer resolution (Wagner, 2018). The lower resolution should make it difficult or

impossible to measure the timing of memory reads to an accurate enough degree for a side channel attack to function. In other words, this does not stop Spectre itself but makes it impossible to retrieve the value of the leaked secret from the microarchitecture.

Another solution that has been developed to mitigate Spectre is known as a “return trampoline” or “retpoline” (Intel, June 2018). This is a method of steering the CPU in an infinite loop in the transient domain to stop it from speculating on the target of an indirect jump, achieving the same effect as turning off speculation for specific portions of a victim program. This means that speculation operates normally, but vulnerable pieces of code can be surrounded by a retpoline to prevent them from executing in the transient domain and leaking data.

The other main mitigation recommended by CPU manufacturers for Spectre is to add an “LFENCE” instruction before critical pieces of code (AMD, 2020; Intel, Jan. 2018). This instruction does not execute until all previous instructions have finished executing, stopping the critical code from executing speculatively and keeping Spectre from exploiting it.

4.3 ZombieLoad

Zombieload (Schwarz et al, 2019) is another type of transient execution attack that works by leaking recently loaded or stale values across logical cores using the CPU’s fill buffer. This attack uses “zombie loads” or more specifically, loads of data that fault at the architectural or microarchitectural level and need to be re-done at a later time. These loads can result in the “dead” data being loaded transiently, aka a zombie load. Once the data has been transiently loaded, it is retrieved using a side channel attack. This attack can currently bypass mitigations for Spectre and Meltdown because it uses the CPU fill buffer instead of the cache. Currently, there are no low performance impact protections that work to fully mitigate ZombieLoad. The high-performance impact protection is to completely disable hyperthreading and flushing the microarchitectural states during context switches. Note that variance in CPU microarchitecture design and features results in ZombieLoad attacks working intermittently across varying CPU architectures.

5. Experience

Throughout our experience evaluating transient execution side channel attacks, we encountered many oddities that make researching and troubleshooting them difficult. The first of these difficulties was the fact that much of the information that exists on these topics is outdated, contradictory, or otherwise difficult to parse and understand. The second and most prevalent of these difficulties was instantiating proof-of-concept codebases we found for these attacks. Many of the codebases we came across either did not work or worked only partially, and it was common for them to have little accompanying documentation with the code base. Compounding the issue, the accuracy and functionality of much of this code is inconsistent across devices. What works flawlessly on most devices may not work at all on another. This, combined with the limited knowledge available about in-depth CPU architecture due largely to the proprietary nature of the information, makes troubleshooting these proofs-of-concept non-trivial.

5.1 Transient Execution Practicality Issues

These transient execution attacks each seemed to have their own issues when it comes to their practical use in a real-world attack in our experience. Of the different attacks we researched and experimented with, Meltdown seemed to have the fewest limitations and the highest potential for use in an actual attack before the patches that mitigated it’s usefulness, whereas Spectre and ZombieLoad both have their own set of limitations.

5.1.1 Spectre Limitations

We came across several limitations that make Spectre impractical or inefficient. One of the main issues we had with Spectre was with consistency, namely being able to get the same code to work on different systems in a consistent manner. The main issues with this appeared to revolve around using either the flush and reload or prime and probe side channel. Spectre implementations using flush and reload worked with 100% accuracy on 2 of the 3 systems we had access to for testing, but not at all on the third. Conversely, prime and probe Spectre implementations we tried worked only partially on the third system (with very low and inconsistent accuracy), but not at all on the first 2. This inconsistent accuracy is illustrated in Table 1, where we have tested all currently available GitHub repositories implementing Spectre and recorded the results per machine. The first and second of the 3 machines tested gave identical results, so they have been grouped together. The third machine uses a pre-release processor from Intel with new, largely undisclosed security features built in, likely contributing

heavily to the differences in attack performance between it and machines 1 and 2. Interestingly, this machine did not defeat all of the proofs-of-concept despite its higher security configuration.

Table 1: Spectre repositories were tested on 3 different machines running Linux, the first of which was virtualized on an Intel® Core™ i7-6700HQ, the second of which was running natively on an Intel® Core™ i5-8259U, and the third of which was running natively on an unidentified Intel® Xeon™ Ice Lake processor

✓ = works, ✗ = does not work, ~ = works partially, - = does not build	Machine 1 & 2	Machine 3
Anonymous, spectre_poc.c, (2018), GitHub repository, https://gist.github.com/anonymous/99a72c9c1003f8ae0707b4927ec1bd8a	✓	✗
Eugnis, spectre-attack, (2018), GitHub repository, https://github.com/Eugnis/spectre-attack	✓	✗
oopsxq, exploit-cve-2017-5715, (2018), GitHub repository, https://github.com/opsxq/exploit-cve-2017-5715	✓	✗
idea4good, spectre, (2018), GitHub repository, https://github.com/idea4good/spectre	✗	✗
fixwu, spectre-attack-demo, (2018), GitHub repository, https://github.com/fixwu/spectre-attack-demo	✓	✗
ogvwzq, spectre, (2019), GitHub repository, https://github.com/ogvwzq/spectre	✗	✗
tbodt, spectre, (2019), GitHub repository, https://github.com/tbodt/spectre	✗	✗
asm, deep_spectre, (2018), GitHub repository, https://github.com/asm/deep_spectre	-	-
ssstonebraker, meltdown_spectre, (2018), GitHub repository, https://github.com/ssstonebraker/meltdown_spectre	✓	~
nsbits, spectre, (2021), GitHub repository, https://github.com/nsbits/spectre	✗	✗
crozone, SpectrePoC, (2019), GitHub repository, https://github.com/crozone/SpectrePoC	✓	✗
amosbe, spectre-without-shared-memory, (2019), GitHub repository, https://github.com/amosbe/spectre-without-shared-memory	✗	✗
chaitanyarahalkar, Spectre-PoC, (2019), GitHub repository, https://github.com/chaitanyarahalkar/Spectre-PoC	✓	✗
LunarResearch, Spectreprime-attack, (2018), GitHub repository, https://github.com/LunarResearch/Spectreprime-attack	-	-
ixtal23, spectreScope, (2018), GitHub repository, https://github.com/ixtal23/spectreScope	✓	✗
mmxrup, CVE-2018-3639, (2019), GitHub repository, https://github.com/mmxrup/CVE-2018-3639	✓	✓
kianenigma, meltdown-spectre, (2019), GitHub repository, https://github.com/kianenigma/meltdown-spectre	✗	✗
charmerkai, spectreRSB, (2020), GitHub repository, https://github.com/charmerkai/spectreRSB-POC	✗	✗
Anton-Cao, spectrev2-poc, (2020), GitHub repository, https://github.com/Anton-Cao/spectrev2-poc	✗	✗
soso7885, spectre-atk, (2018), GitHub repository, https://github.com/soso7885/spectre-atk	✓	✗
vaamonde, spectre-meltdown, (2018), GitHub repository, https://github.com/vaamonde/spectre-meltdown	✓	✗
Hackndo, spectre-poc, (2019), GitHub repository, https://github.com/Hackndo/spectre-poc	✓	✓
cristianomello, spectre, (2018), GitHub repository, https://github.com/cristianomello/spectre	✓	✗
WPC-CS830-Spectre, Spectre1 PoC, (2019), GitHub repository, https://github.com/WPC-CS830-Spectre/Spectre1_PoC	~	✗
autoint, SpectreVariant1, (2019), GitHub repository, https://github.com/autoint/SpectreVariant1	✗	✗
WPC-CS830-Spectre, Spectre2 PoC, (2019), GitHub repository, https://github.com/WPC-CS830-Spectre/Spectre2_PoC	✗	✗
autoint, SpectreVariant4, (2019), GitHub repository, https://github.com/autoint/SpectreVariant4	-	-
Markus-MS, Spectre-Attack, (2019), GitHub repository, https://github.com/Markus-MS/Spectre-Attack	✓	✓
jottama, SpectrePoC, (2018), GitHub repository, https://github.com/jottama/SpectrePoC	✓	✗
qutorial, spectre, (2018), GitHub repository, https://github.com/qutorial/spectre	-	-
srishanbhattarai, spectre, (2019), GitHub repository, https://github.com/srishanbhattarai/spectre	-	-
andrey-martynyuk, Spectre, (2019), GitHub repository, https://github.com/andrey-martynyuk/Spectre	✓	✗
Bookvar, spectre, (2021), GitHub repository, https://github.com/Bookvar/spectre	✓	✗

Furthermore, we found that not only did the two different side channels used not work across different systems consistently, but they also did not seem to hold up to their theoretical abilities. The best example of this is with flush and reload, which should be able to work across physical cores by operating in the last level cache, but we found that this was not the case. Over the course of our testing, flush and reload based attacks always needed

the attacker and victim to be pinned to the same core as the side channel attack would only succeed in the L1 or L2 cache.

The second issue we found is that even if the attack does work on the hardware it's being run on, the attacker must have a significant amount of control over the victim program in order to steal data from it. It needs to be able to call functions within the victim with specific, attacker-controlled inputs. In most cases the amount of access, permissions, and knowledge required would make it more efficient to simply use a different privilege escalation attack to achieve the same result.

5.1.2 ZombieLoad Limitations

ZombieLoad, while an interesting concept and even while being reasonably fast and accurate in the values it is reading, is again not practical for most attacks. This stems largely from the fact that the attacker does not get a choice of what data ZombieLoad recovers as it simply grabs whatever was in the last load. This could be useful as an attack in exactly the right conditions, but in most scenarios would require the same data to be accessed many times in a row to steal any relevant victim secrets. An attack such as leaking a single AES might take years to accomplish. This means that a different type of attack will usually be a better option.

5.2 SpectreXP: A Cross-Process Spectre Proof-of-Concept

One issue we found when researching Spectre and attempting to design a practical attack that could simulate a real-world data leak was that most of the proof of concepts that currently exist are only a single process, which steals data from inside itself by reading out of bounds from an array. We were unable to get the ones that do implement Spectre across multiple processes working for various reasons, mostly relating to issues with the side channels being used.

To prove that Spectre can in fact work across multiple processes, we created a novel Spectre proof-of-concept, which we have dubbed as "SpectreXP". Our proof-of-concept consists of a simple victim program that is intentionally vulnerable to Spectre by containing code like that shown in Figure 5. The attacker program sends inputs to the victim via a socket. As with other Spectre variants, the attacker sends many valid values as inputs to the victim program to mistrain the CPU branch predictor. Once the branch predictor has been sufficiently trained, the attacker sends the victim an input that is invalid. This will cause the victim to speculatively access secret memory, leaving changes in the cache that the attacker can recover, as shown in Figure 6.

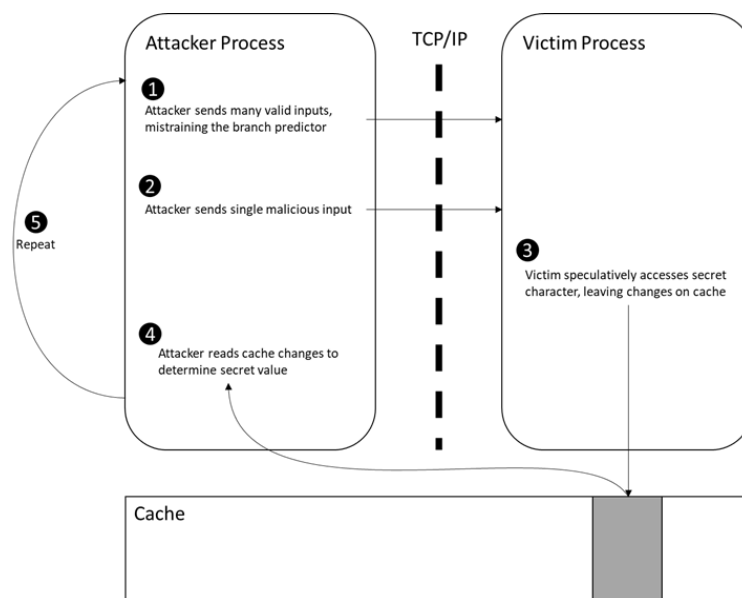


Figure 6: SpectreXP: 1 – the attacker trains the branch predictor with valid data. 2 - the attacker sends a malicious input to force a speculative read. 3 – the victim uses that input to read arbitrary data speculatively. 4 – the attacker reads microarchitectural changes using flush and reload to determine part of secret value. 5 – repeat until entire secret has been read.

This proves that it is possible to get Spectre to work across multiple processes as long as they share memory in the form of the cache, specifically the L1 or L2 caches due to the found limitations of flush and reload. This means

that the 2 processes also need to be pinned to adjacent logical cores via hyperthreading. The code for this new proof of concept can be found here: <https://github.com/riversideresearch/spectreXP>

6. Conclusion

In this paper, we presented our research into timing-based side channel attacks and the Meltdown, Spectre, and ZombieLoad attacks, as well as going over our experiences with experimenting with these types of attacks on real hardware. We found that the theoretical properties of side channel attacks did not always hold up on real hardware and were not consistent across different hardware, meaning that the attacks that rely on them also often fail to work partially or entirely. Side channels are also difficult to research due to the outdated, contradictory, or complex nature of much of the existing research. We also learned that debugging these experiments can often prove to be non-trivial due to the proprietary, poorly documented, or otherwise non-disclosed nature of the hardware or software being worked with. Due to these issues, we have come to the conclusion that there are few, if any, situations in which transient execution attacks are truly effective or efficient against current technology in the real world.

6.1 Future Work

Future work would revolve around gathering up all the information available on this topic and verifying it piece by piece. The goal of this would be to build a “unified theory” of side channel and transient execution attacks, finding and correcting as many inaccuracies and contradictions as possible while compiling and standardizing the ideas, vocabulary, and practices of existing work. This would allow us to dig deeper into the intricacies of CPU architecture and learn why, for instance, certain attacks do not work consistently across different hardware. This goal would benefit greatly from the testing of these attacks on more machines and architectures than we had access to for experimentation purposes, such as those running AMD or ARM processors. Other metrics that would be useful to record about these exploits in the course of this research would be the accuracy in retrieving known data over many trials, how quickly they receive that data, and how easy they are to detect and mitigate. Learning about these topics and collecting data from additional experiments would then allow us to create faster, more accurate, and more resilient attacks to push the boundaries of these technologies.

One “final” goal to come out of all this would be that of building a proof-of-concept for an attack that could be dropped onto nearly any combination of hardware and software and run with nearly 100% accuracy, finally allowing us to circumvent the need for heavy customization per each system being attacked. Another possible goal is using transient execution and side channels to break the barrier between host and guest in a virtualized environment and wrongly pass data back and forth from guest to host or vice versa.

References

- AMD. (2020, September) Software Techniques for Managing Speculation on AMD Processors.
- Canella, C., Genkin, D., Giner, L., Gruss, D., Lipp, M., Minkin, M., Moghimi, D., Piessens, F., Schwarz, M., Sunar, B. and Van Bulck, J. (2019, November) “Fallout: Leaking data on meltdown-resistant cpus”, Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 769-784).
- Canella, C., Khasawneh, K.N. and Gruss, D. (2020, September) “The evolution of transient-execution attacks”, Proceedings of the 2020 on Great Lakes Symposium on VLSI (pp. 163-168).
- Canella, C., Van Bulck, J., Schwarz, M., Lipp, M., Von Berg, B., Ortner, P., Piessens, F., Evtushkin, D. and Gruss, D. (2019) “A systematic evaluation of transient execution attacks and defenses”, 28th USENIX Security Symposium (USENIX Security 19) (pp. 249-266).
- Godbolt, Matt. (2018, 13 January) “Meltdown and Spectre.” YouTube, uploaded by Matt Godbolt, [online], <https://youtu.be/IPhVL3A-e6E>
- Gregg, Brendan. (2018, 9 February). “KPTI/KAISER Meltdown Initial Performance Regressions.” [online] <https://www.brendangregg.com/blog/2018-02-09/kpti-kaiser-meltdown-performance.html>
- Gruss, D., Lipp, M., Schwarz, M., Fellner, R., Maurice, C. and Mangard, S. (2017, July) “Kaslr is dead: long live kaslr”, International Symposium on Engineering Secure Software and Systems (pp. 161-176). Springer, Cham.
- Intel. (2018, January) Intel Analysis of Speculative Execution Side Channels.
- Intel. (2018, June) “Retpoline: A Branch Target Injection Mitigation”, [online], <https://software.intel.com/content/dam/develop/external/us/en/documents/retpoline-a-branch-target-injection-mitigation.pdf>
- Kernel.org. “Page Table Isolation”, [online] <https://www.kernel.org/doc/html/latest/x86/pti.html>
- Kocher, P., Horn, J., Fogh, A., Genkin, D., Gruss, D., Haas, W., Hamburg, M., Lipp, M., Mangard, S., Prescher, T. and Schwarz, M. (2019, May) “Spectre attacks: Exploiting speculative execution”, 2019 IEEE Symposium on Security and Privacy (SP) (pp. 1-19). IEEE.

- Lipp, M., Schwarz, M., Gruss, D., Prescher, T., Haas, W., Fogh, A., Horn, J., Mangard, S., Kocher, P., Genkin, D. and Yarom, Y. (2018) "Meltdown: Reading kernel memory from user space", 27th USENIX Security Symposium (USENIX Security 18) (pp. 973-990).
- Liu, F., Yarom, Y., Ge, Q., Heiser, G. and Lee, R.B. (2015, May), "Last-level cache side-channel attacks are practical", 2015 IEEE symposium on security and privacy (pp. 605-622). IEEE.
- McIlroy, R., Sevcik, J., Tebbi, T., Titzer, B.L. and Verwaest, T. (2019) "Spectre is here to stay: An analysis of side-channels and speculative execution", arXiv preprint arXiv:1902.05178.
- Noack, L. and Reichert, T. (2018) "Exploiting Speculative Execution (Spectre) via JavaScript", Advanced Microkernel Operating Systems, p.11.
- Osvik, D.A., Shamir, A. and Tromer, E. (2006, February) "Cache attacks and countermeasures: the case of AES", Cryptographers' track at the RSA conference (pp. 1-20). Springer, Berlin, Heidelberg.
- Schwarz, M., Lipp, M., Moghimi, D., Van Bulck, J., Stecklina, J., Prescher, T. and Gruss, D. (2019, November) "ZombieLoad: Cross-privilege-boundary data sampling", Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 753-768).
- Tromer, E., Osvik, D.A. and Shamir, A. (2010) "Efficient cache attacks on AES, and countermeasures", Journal of Cryptology, 23(1), pp.37-71.
- Van Bulck, J., Minkin, M., Weisse, O., Genkin, D., Kasikci, B., Piessens, F., Silberstein, M., Wenisch, T.F., Yarom, Y. and Strackx, R. (2018) "Foreshadow: Extracting the keys to the intel SGX kingdom with transient out-of-order execution", 27th USENIX Security Symposium (USENIX Security 18) (pp. 991-1008).
- Van Schaik, S., Milburn, A., Österlund, S., Frigo, P., Maisuradze, G., Razavi, K., Bos, H. and Giuffrida, C. (2019, May) "RIDL: Rogue in-flight data load", 2019 IEEE Symposium on Security and Privacy (SP) (pp. 88-105). IEEE.
- Wagner, L. (2018, 3 January) "Mitigations landing for new class of timing attack", [online], <https://blog.mozilla.org/security/2018/01/03/mitigations-landing-new-class-timing-attack/>
- Weisse, O., Van Bulck, J., Minkin, M., Genkin, D., Kasikci, B., Piessens, F., Silberstein, M., Strackx, R., Wenisch, T.F. and Yarom, Y. (2018) "Foreshadow-NG: Breaking the virtual memory abstraction with transient out-of-order execution".
- Xiong, W. and Szefer, J. (2020) "Survey of transient execution attacks." arXiv preprint arXiv:2005.13435.
- Yarom, Y. and Falkner, K. (2014) "FLUSH+ RELOAD: A high resolution, low noise, L3 cache side-channel attack", 23rd USENIX Security Symposium (USENIX Security 14) (pp. 719-732).

Bug Bounties: Between New Regulations and Geopolitical Dynamics

Jantje Silomon, Mischa Hansel and Fabiola Schwarz

Institute for Peace Research and Security Policy, Hamburg, Germany

silomon@ifsh.de

hansel@ifsh.de

schwarz@ifsh.de

Abstract: Crowdsourced security and vulnerability co-ordination platforms, such as Bugcrowd or HackerOne, reward individuals for discovering, reporting, and responsibly disclosing software bugs. A growing number of vendors are turning towards these platforms to improve their product's security, whilst others set up their own bug bounty programs (BBPs) alongside more traditional approaches, such as in-house testing and professional security reviews. Whether providing a supplementary or even alternative path to organisational cybersecurity, these newer approaches go beyond increasing product security, for example by fostering co-operation between various actors or providing a clear incentive to remain on the ethical side of security research. Whilst some research centres on the reward structures, actor motivations, or effectiveness, the wider impact on peace and stability in cyberspace is rarely examined. Similarly, rarely is light shed on emerging regulatory or policy approaches, or the effects this might have. To fill these gaps, the paper will use Global Public Goods (GPGs) theory to example BBPs across two case studies. Whereas the novel Chinese regulations push towards more national sovereignty in cyberspace, the European Union invests in the compensation of BBP under-provision among open source software (OSS). These regulatory changes in China and endeavours by the European Union, respectively, reveal that the prevalent geopolitical divisions in related topics, such as internet governance, continue to play their part. Further research on BBPs is proposed to quantitatively examine their effect on peace and stability.

Keywords: Bug Bounty Programs, Vulnerability Disclosures, Global Public Goods, Geopolitical Dynamics

1. Introduction

As weak points are an inherent part of computer systems, various forms of vulnerability management have emerged over the years, ranging from best practices to responsible disclosure programs. Processes have become more formalised and security researchers are no longer left in the dark, having “to identify potential buyers, to determine accurate prices, and to overcome various barriers in executing a sale” (Kuehn, 2018). One example are BBPs, either hosted on platforms or created by vendors themselves, open to the public or by invitation only. These reward individuals for discovering and reporting software bugs, with the pay-out usually depending on the severity of the vulnerability (Malladi and Subramanian, 2019). There are also other intermediaries, some of which have come under scrutiny for also selling to disreputable entities (Perloth, 2021; Coble, 2020).

Most BBPs set out specific criteria, including scope, submission and review processes, time constraints, as well as terms and conditions. These programs might focus on a single product, an element thereof, encompass an entire product suite or vulnerability type. Vulnerability information usually becomes public in line with responsible disclosure processes by the companies and platforms. Until recently, BBPs have remained rather apolitical. However, with growing success and publicity, there is also a rise in political alertness. Public institutions, such as the US Pentagon (Chatfield and Reddick, 2017) or the EU, have started to offer BBPs to secure their own networks or enhance OSS security, respectively. Questions regarding responsibility have arisen, from debates on product lifecycle support and its extension to whether or not vulnerability information is a public good. The latter in particular is of interest here in light of new regulations having come into effect in China as of September 2021 (CAC, 2021).

Against this backdrop, the paper will first delve into GPG theory before arguing why aspects of BBPs should be treated as such. Next, two empirical case studies with different focal points will be analysed, beginning with China and followed by the European Union. The former centres on new regulatory changes and extrapolates how these might affect BBPs, and cyber security more broadly. The latter addresses a specific programme that attempts to deal with OSS security. A preliminary model is then formed and discussed prior to concluding remarks.

2. Global Public Goods Theory

A public good combines two qualities, namely non-rivalry and non-excludability. Non-rivalry means that consumption of the good by one actor does not result in a reduced availability of said good for others. Non-excludability on the other hand refers to the practical impossibility of excluding any actors from consuming the

good (Samuelson, 1954). If benefits are shared on a global scale, public goods essentially constitute GPGs. The use of public goods usually leads to either positive or negative externalities, potentially again of a global reach.

Due to their non-rivalry and non-excludability, pure public goods face two caveats leading to under-provision (Kaul et al, 1999). The first problem is free riding, as free access to the good allows actors to benefit without contributing to availability (Boonen et al, 2019). Consequently, there are no incentives to contribute in the first place. Under these circumstances, the provision of public goods may become insufficient or even break down completely. The second is known as the prisoner's dilemma, a game theoretical paradox that adds uncertainty about the behaviour of others to the picture. In the view of incomplete knowledge, rational actors must decide whether to trust the other party and act responsibly or not. While each player prefers mutual trust and cooperation to mutual defection, the best outcome for each player is a situation in which he defects while the other side chooses to cooperate. In turn, this means that cooperating while the other player defects results in the worst possible outcome (Hansel et al, 2018). Consequently, and despite common interest, rational actors are tempted to act uncooperatively, for only by cheating can exploitation by others be avoided.

2.1 Bug Bounties and GPGs

The question remains whether cybersecurity, or more specifically BBP-generated vulnerability information, can be regarded as a GPG. Following Taddeo (2019), a system's robustness would fall under the definition, resilience and responses to attacks would not. Vulnerability information is, among other things, part of a robust system (Taddeo 2019), with Rosenzweig (2012) sharing the view and calling cybersecurity information a public good. This contrasts with commercial IT security solutions, such as anti-virus software or intrusion-detection systems, which are bought and sold on the private market and are thus excludable. The information about vulnerabilities, however, is not. Moreover, it fulfils the idea of non-rivalry as sharing with others does not reduce its usefulness to each collaborating partner.

While software vendors pay for BBPs, customers do so at most indirectly by assuming that part of the price paid for the product service is its (security) maintenance. In this way, everyone worldwide employing a specific software is benefitting from that vendor's BBP and other measures. This can be seen as a positive externality from an economic perspective. Rationally thinking, if competitor A on the market invests in secure products, it will likely attract more customers in the long run, as it develops a more sustainable product. This may lead competitor B and C to similarly invest in product security (Böhme, 2006). In view of potential misuses of computers, for example via botnets, security should rather be centred on a node as opposed to the network level (ibid.). The weakest node is likely the primary point of attack, acting as a floodgate to the whole system or network, and can therefore lead to a negative externality. In turn, a positive externality would be if one node invests in security. This way, it automatically reduces the likelihood for others to be attacked via that node's previously closed gates.

Hence, although investing in BBPs might only directly benefit own customers, it has a positive spillover effect on users of other software, too. This indirect benefit qualifies the information on vulnerabilities (resulting from BBPs) for being non-excludable while at the same time being non-rival. However, and following Weber (2017), it ultimately depends on policy choices whether vulnerability information is shared or retained by companies and governmental agencies. Stemming from this, different countries choose different strategies in governing vulnerability information.

3. Empirical Case Studies

In the following, two empirical case studies serve the purpose of inductively exploring real-world variance of BBP policies. Case selection was driven by data availability and the benefit of analysing most unlike regulatory frameworks, i.e. a focus on recent European and Chinese public policy decisions. At the same, they are of considerable intrinsic importance, given the respective market sizes.

3.1 China

On June 1, 2017, the "Cybersecurity Law of the People's Republic of China" (中华人民共和国网络安全法), commonly referred to as the Chinese Cybersecurity Law, came into effect (Creemers, Triolo, and Webster, 2018). Its Article 26 states that the publishing of vulnerabilities and other cybersecurity information "shall comply with relevant national provisions" (ibid.) and there are signs that China has been exerting pressure on its security researchers. International concerns grew when China announced not to send any security researchers to the Canadian

hacking competition ‘Pwn2Own’ in March 2018 (Krahulcova, 2018). A year later, in 2019, the first Chinese vulnerability disclosure regulation draft was released, requiring vulnerabilities to be reported exclusively to state authorities in a first step (Udemans, 2019). In July 2021, the Cyberspace Administration of China (CAC) issued the “Regulations on the Management of Network Product Security Vulnerability” (网络产品安全漏洞管理规定) coming into force only two months later (CAC, 2021). While its Article 7 encourages the creation of BBPs, it also establishes the state as the dominant gatekeeper on vulnerability information (Cimpanu, 2021). Thus, vendors are obligated to notify the Ministry of Industry and Information Technology (工业和信息化部) within 48 hours regarding any reported vulnerabilities. Article 9 goes even further, essentially preventing any Chinese security researchers from cooperating with (non-vendor) BBPs as they are no longer allowed to disclose vulnerability information to “overseas organizations or individuals other than network product providers” (ibid.).

These new regulations and decisions by the Chinese government send a twofold signal. First, it seems that China is free riding on the efforts of international BBPs. This does not mean that there are no Chinese efforts in setting up own BBPs or hacking competitions, as the examples of Huawei (Whittacker, 2019) or the 2021 Tianfu Cup show. Chinese BBPs make use of the global crowd of security researchers to secure their own software products. At the same time, the government does not allow its own citizens to participate in foreign BBPs. This will have a significant impact on the BBP market, according to HackerOne (2020), as Chinese security researchers are second in reporting vulnerabilities in terms of bounty in-flow.

The second and even more important element arises from the previously discussed prisoner’s dilemma innate to the governance of GPGs. Given that an increasing number of states not only invest in defensively orientated security measures but also in offensive capabilities (Herpig, 2018), state security agencies have an interest in exploiting vulnerability information instead of sharing them with third parties. Here, Zero-Day Exploits (ZDE) are particularly sought after, which refer to a vulnerability that is yet unknown to the developer or vendor but exists ‘in the wild’. China appears to be leaning towards this direction, having also assigned ZDEs a strategic value (O’Neill, 2021; Hewage and Ukwandu, 2021). The stockpiling of ZDEs results in even greater distrust, in turn acting as a motivator for other actors to also withhold information.

Drawing on the simplified (two-player) representation of the prisoner’s dilemma, the situation could be depicted as in the table below. The most desirable outcome for international security would be a BBP market open to all vendors and security researchers where gained vulnerability information is disclosed responsibly to all parties (top left). However, recent signals sent from China indicate that it decided to put national security interests first (Thorne and Hoffman, 2021) and to move vulnerability disclosure from a GPG towards a club good, only available to vendors and the state within its borders.

		Actor A	
		Disclosure	Retention
Actor B	Disclosure	Responsible vulnerability disclosure in both states	Responsible vulnerability disclosure in State B, retention thereof in state A
	Retention	Responsible vulnerability disclosure in State A, retention thereof in state B	Both states withhold vulnerability information

One could now argue that there is nothing new about this dynamic because numerous governments around the world withhold vulnerabilities under certain circumstances, and the relevant procedures and criteria for this are also evident in many cases through so-called Vulnerabilities Equities Processes (VEPs). There is, however, a crucial difference between for example the US VEPs and the new Chinese regulations. While VEPs only decide on vulnerabilities discovered by government agencies, the Chinese government also secures access to those by third parties. Although the information is due to be shared with manufacturers as well, state intelligence services

or other government actors will have a comfortable window of opportunity to exploit reported vulnerabilities. The reason for this is the average time required for the development, testing, and distribution of patches before any effective defence can be in place. To underline this argument, consider the difference between the 30, 60 or even 90 days granted to software vendors by most BBPs prior to details being publicly released, and the maximum 48-hour time lag after which reports have to be shared with the Chinese government. Arguably, this leaves at least almost a month for exploits to be developed and used by Chinese intelligence, military, or other agencies.

Against this backdrop, Chinese BBP nationalisation and the associated withholding of vulnerability information has serious implications for international security and governance that go beyond the programs as such. On the one hand, the nationalisation of white hats will likely be seen as a threatening move to expand China's 'cyber arms arsenal', which might lead to no fewer destabilising countermeasures. On the other hand, forcing Chinese security researchers to withdraw from transnational BBPs arguably undermines the effectiveness of a transnational bottom-up mechanism of 'proliferation control' in cyberspace (Silomon, 2020).

3.2 The European Union

Popularity of OSS has skyrocketed, among not only public authorities and businesses but also individuals. In their main form, OSS licenses allow software to be freely used, modified, and shared. In many cases, development is often community-based (Berlinguer, 2020). While the responsibility to update and further develop commercial software lies with the vendor, OSS maintenance responsibility is not as clearly assigned. Often security measures including post-release bug hunting, reporting, and/or patching, are left to individuals or groups on a voluntary basis lacking financial reward structures. Given OSS development nature and often lack of revenue, it is not surprising that resources are scarce (Rogowsky, 2014). For the same reason, OSS projects cannot compete with commercial software vendors on the BBP market.

Under-protection therefore represents a classical example of market failure, where lack of incentives to invest in security leads to under-provision of (OSS-specific) vulnerability information as a GPG. While a number of private initiatives such as the Core Infrastructure Initiative and Google's Project Zero seek to address this problem, the EU launched its own hands-on solution in 2015. Its Free and Open Software Auditing (EU-FOSSA) project started with an initial one million Euro budget (European Commission, 2020) focussing on auditing two critical software programmes and running a cybersecurity awareness campaign. After its completion, EU-FOSSA 2 was launched with a wider scope in 2019. This iteration encompassed a greater number of European institutions, research on OSS use among public administrations worldwide, and, most importantly, fifteen BBPs as well as three hackathons. A core aim was to mitigate the under-provision of OSS security by creating a BBP that could match the rewards offered by commercial BBPs on the market.

Nonetheless, there is also an industrial policy rational for the EU's efforts as OSS can guarantee independence from foreign commercial software providers. Furthermore, publicly financed BBPs could support the entrance of European competitors to the current US dominance on the global BBP market (HackerOne, 2020). There have been calls for Europe to catch up and act as a counterbalance (Schulze, 2019), seeking to encourage more European-based firms to invest in crowdsourced security. In turn, this could improve European sovereignty in terms of vulnerability information and prevent foreign intelligence services gaining access or misusing the information. YesWeHack, founded in 2013, serves as a good example for a growing European stake, having grown by 450% in venture capital funding between 2019 and 2021, while security researcher participation has risen by 350% (YesWeHack, 2019 and 2021).

4. A Preliminary Theory of Regulatory Models

Both case studies illuminate ways of complementing, if not replacing, the existing primarily market-driven BBP environment. In very simplified terms, this results in a total of three models, or ideal types, for BBP and security research regulation, thus mirroring the development within other areas of internet and cybersecurity governance, such as data privacy (Aho and Duffield, 2020; Siebert, 2021). In the following section, we further elaborate on these models by offering a theoretically informed comparison of their guiding principles, diffusion potential, and impact. Starting with principles and norms, an ideal-type comparison will delve into the following characteristics:

1. The first model, emerging within the Chinese context, follows the principles of digital authoritarianism and absolute state sovereignty, and applies them to the governance of BBPs. National governments,

accordingly, have a right and duty to effectively control and capture any security research on ICT vulnerabilities within their jurisdiction, including the enabling intellectual and material resources.

2. The second model is diametrically opposed to the first, trusting essentially the market with almost all governance issues, especially as they relate to allocation of resources and the distribution of benefits. The role of the government instead is limited to making sure that an enabling legal framework is in place, for example regarding the protection of security researchers from arbitrary claims for damages. Beyond that, the government uses BBPs only for the security of its own systems.
3. Finally, European public funding of BBP on OSS may well be regarded as the nucleus of a paradigmatic middle ground, where a majority of BBPs depends on the market, but public spending is used to strategically correct for market failures and to ensure that the benefits of vulnerability research are as equally distributed as possible.

Currently, the impact of each of these models on peace and security is undoubtedly speculative, and only practical experience will show whether certain opportunities or risks materialise. Nonetheless, it is likely that the second model will result in at least some under-provision of GPGs, neither preventing free riding and thus insufficient resource allocation, nor safeguarding against misallocation, i.e. a mismatch between funding priorities and actual areas of particular concern. Furthermore, critics point out that some companies use BBPs to pressure security researchers into non-disclosure-agreements. Others use these platforms as a cheap and less demanding alternative to setting up proper vulnerability disclosure processes themselves (Porup, 2020). In both cases, use of BBPs essentially prevents rather than produces the disclosure of vulnerability information as GPG. The European effort, in contrast, while being capable of closing crucial protection gaps at the regional level, does not provide a solution to the equally challenging use of unlicensed, and therefore mostly unpatched, software. To maximise the European model, it is therefore vital to complement it with measures promoting OSS use in many countries.

Assessing the impact on peace and security of the first model, in comparison, appears straightforward. After all, the transformation of BBPs, leading to the shift of vulnerability information from a global public into a club good, with state authorities as the central gatekeeper, is certainly regrettable by systematically taking away protection from internet users and organisations worldwide. Even if there is no or only limited intention to use vulnerabilities, the mere knowledge of the nationalisation of vulnerability research exacerbates the international security dilemma and will thus probably fuel the ongoing militarisation of cyberspace. Even if the Chinese government was to publicly declare that reported security vulnerabilities would not be used for state operations, this would not change. The main reason is that this cannot be verified externally, exemplifying structural dynamics of the prisoner's dilemma discussed above.

Whatever their impact on peace and security, emerging models of regulating BBPs will matter even more when they diffuse beyond national borders, through emulation or by sanctioning others for not adopting such standards. As has been already indicated, these models mirror competing governance paradigms within related policy areas in many ways, such as the regulation of free speech or data privacy. However, as these rivalries have progressed further, they might hold some insights that could be applied to the case of BBPs and vulnerability disclosure regulation. For example, it has often been said that the "Brussels effect" (Bradford, 2021), i.e. the size of the European single market amongst other issues, is the reason that EU data protection policies are influential at the global level despite the dominance of US, and increasingly Chinese, internet companies. Thus, being able to deny or provide access to costumers can outweigh the lack of production capacities in this particular regulatory competition. Yet here, the case might be different as customers in this context are not European internet users, but the internet companies and software vendors themselves. Nevertheless, the possibility remains that others will follow the European model because of its perceived benefits or legitimacy in comparison to alternative forms.

A final characteristic of competition is that internet governance paradigms can only coexist with each other to a limited extent, as they each generate extraterritorial effects that cause conflicts at the international level (Internet Society, 2018). Governance of BBPs is likely no exception in this regard, and one such effect has already been mentioned, i.e. a reduction of available human resources within Western BBPs as the result of the forced withdrawal of Chinese security researchers. There are likely many others and additional research is needed to be able to anticipate such effects, which in themselves can be drivers of international conflict and mistrust.

5. Conclusion

Based on the growing financial volume and the increasing prevalence of BBPs worldwide used to disclose information on vulnerabilities, our paper focused on the question to what extent such bottom-up mechanisms can be understood as a contribution to global cybersecurity and thus to peace and security in general. Using the theory of GPGs, we differentiated existing practices as well as emergent regulatory models that compete with the non-commercial and apolitical original model. While commercial BBPs already dominate the field, regulatory approaches in China and the EU underline a growing trend of politicisation and state intervention. The future effects of these models on the availability of vulnerability information are foreseeably very different.

However, more research is needed to come to a more reliable conclusion in this regard and to enable adequate policy responses to possible unintended effects. First, we propose systematic quantitative studies to test deductive assumptions about the specific benefits of BBPs for peace and security with robust empirical data. Of particular interest is the extent to which the profile of vulnerabilities exploited in the context of particularly disruptive cyberattacks matches the typical characteristics of vulnerabilities discovered and reported through BBPs. The more plausible such matches are, the more this would suggest that BBPs could indeed make a significant contribution to the prevention of serious cyber incidents. While studies on the overlap between vulnerabilities used by APTs and information included within state-provided vulnerability databases in Russia and China already exist (see Leyden, 2018), the authors of this paper themselves are currently preparing a comprehensive research design particularly on BBPs.

Second, future qualitative studies could trace the relative diffusion of these or other governance models to different world regions and by various mechanisms. The latter can be processes of voluntary imitation as well as incentive-based adoption of standards. Finally, the role of transnational processes of norm- and standard-setting in the context of intergovernmental, but also private-sectoral cooperation, such as an obligation for vendors to offer BBPs (Schulze and Reinhold, 2018) should be discussed. Overall, there is a huge potential for research on various aspects of emergent BBP governance models, ranging from underlying problems to the significance of their effects on global peace and security and their relative competitiveness in a world of diverging standards and geopolitical divisions.

References

- Aho, Brett and Roberta Duffield (2020) Beyond Surveillance Capitalism: Privacy, Regulation and Big Data in Europe and China, *Economy and Society* 49(2), 187-212.
- Berlinguer, Marco (2020) Commons, Markets and Public Policy: Free and Open Source Software as a Laboratory for the Information Paradigm, Transform! ePaper, January 2020. Retrieved from: https://www.transform-network.net/fileadmin/user_upload/2020-01-commons_3.pdf (accessed: 27th August 2021).
- Böhme, Rainer (2006) A Comparison of Market Approaches to Software Vulnerability Disclosure. In Günter Müller (ed.) *Emerging Trends in Information and Communication Security*, Freiburg: Springer, 298-311.
- Boonen, Christiaan, Nicolás Brando, Samuel Cogolati, Rutger Hagen, Nils Vanstappen and Jan Wouters (2019) Governing as Commons or as Global Public Goods: Two Tales of Power, *International Journal of the Commons* 13(1), 553-577.
- Bradford, Anne (2012) The Brussels Effect, *Northwestern University Law Review* 107(1), 1-68.
- Chatfield, Akemi and Chris Reddick (2017) Cybersecurity Innovation in Government: A Case Study of U.S. Pentagon's Vulnerability Reward Program. In: *Proceedings of the 18th Annual International Conference on Digital Government Research*, States Island, June 2017, 64-73. DOI: 10.1145/3085228.3085233.
- Cimpanu, Catalin (17/07/2021) Chinese Government Lays out new Vulnerability Disclosure Rules, The Record. Retrieved from: <https://therecord.media/chinese-government-lays-out-new-vulnerability-disclosure-rules/> (accessed: 28th August 2021).
- Creemers, Roger, Paul Triolo and Graham Webster (29/06/2018) *Translation: Cybersecurity Law of the People's Republic of China* (Effective June 1, 2017). Retrieved from: <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/> (accessed: 28th August 2021).
- Cybersecurity & Infrastructure Security Agency (CISA) (17/12/2003) Homeland Security Presidential Directive 7: Critical Infrastructure Identification, Prioritization, and Protection. Retrieved from: <https://www.cisa.gov/homeland-security-presidential-directive-7> (accessed: 6th October 2021).
- Cyberspace Administration of China (CAC) 国家互联网信息办公室 (12/07/2021), 工业和信息化部 国家互联网信息办公室 公安部关于印发网络安全漏洞管理规定的通知. 工信部联网安 (2021) 66号. Retrieved from http://www.cac.gov.cn/2021-07/13/c_1627761607640342.htm (accessed: 1st September 2021).
- European Commission (EC) (14/07/2020) *EU-FOSSA 2 - the EU's Open Source Cybersecurity Project Ends*. Retrieved from: https://ec.europa.eu/info/news/eu-fossa-2-eus-open-source-cybersecurity-project-ends-2020-jul-14_en (accessed: 28th August 2021).

- HackerOne (2020) *The 4th Annual Hacker Powered Security Report*. The Study on the Hacker-Powered Security Ecosystem. Retrieved from: <https://www.hackerone.com/hacker-powered-security-report> (accessed: 28th August 2021).
- Hansel, Mischa, Max Mutschler and Marcel Dickow (2018) Taming Cyber Warfare: Lessons from Preventive Arms Control, *Journal of Cyber Policy* 3(1), 44-60.
- Herpig, Sven (2018) Governmental Vulnerability Assessment and Management: Weighing Temporary Retention versus Immediate Disclosure of 0-Day Vulnerabilities, Stiftung Neue Verantwortung e.V., August 2018, Berlin.
- Hewage, Chaminda and Elochukwu Ukwandu (25/05/2021) A Hacking Competition May Have Given China new Ways to Spy on the Uyghurs. Retrieved from: <https://scroll.in/article/995578/a-hacking-competition-may-have-given-china-new-ways-to-spy-on-the-uyghurs> (accessed: 28th August 2021).
- Internet Society (2018) The Internet and Extra-Territorial Effects of Laws, Internet Society Concept Note. Retrieved from <https://www.internetsociety.org/wp-content/uploads/2018/10/The-Internet-and-extra-territorial-application-of-laws-EN.pdf> (accessed: 5th August 2021).
- Kaul, Inge, Isabelle Grunberg and Marc A. Stern (1999) Defining Global Public Goods. In: Inge Kaul, Isabelle Grundberg and Marc A. Stern (eds) *Global Public Goods - International Cooperation in the 21st Century*, Oxford: Oxford University Press, 2-19.
- Krahulcova, Lucie (20/09/2018) China Tips the Scale of Global Cybersecurity by Hoarding Vulnerabilities. Retrieved from: <https://www.accessnow.org/china-tips-the-scale-of-global-cybersecurity-by-hoarding-vulnerabilities/> (accessed: 28th August 2021).
- Kuehn, Andreas (2018) New Paradigms in Securing Software Vulnerabilities - An Institutional Analysis of Emerging Bug Bounty Programs and Their Implications for Cybersecurity, *SSRN Electronic Journal* (1), 1 16. DOI: 10.2139/ssrn.2809862.
- Leyden, John (17/07/2018) Russia's National Vulnerability Database is a bit like the Soviet Union. Sparse and Slow, The Register. Retrieved from: https://www.theregister.com/2018/07/17/russia_vuln_database/ (accessed: 29th August 2021).
- Malladi, Suresh Siva & Hemang Subramanian (2019) Bug Bounty Programs for Cybersecurity: Practices, Issues, and Recommendations, *IEEE Software* 7459(c), 1 9. DOI: 10.1109/MS.2018.2880508.
- Mohajan, Haradhan Kumar (2018) Qualitative Research Methodology in Social Sciences and Related Subjects, *Journal of Economic Development, Environment and People* 7(1), 23-48.
- O'Neill, Patrick Howell (06/05/2021) How China Turned a Prize-Winning iPhone Hack against the Uyghurs. Retrieved from: <https://www.technologyreview.com/2021/05/06/1024621/china-apple-spy-uyghur-hacker-tianfu/> (accessed: 28th August 2021).
- Philippens, Henry (2012) The Future Prospects of De-alerting: Complexities and Considerations for Reducing Risk, *Nuclear Notes* 2(1), 13-24. Retrieved from: https://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/120529_Spies_NuclearNotes2_Web.pdf (accessed: 3rd September 2021).
- Porup, J.M. (2020) Bug Bounty Platforms buy Researcher Silence, Violate Labor Laws, Critics Say, CSO Online. Retrieved from: <https://www.csoonline.com/article/3535888/bug-bounty-platforms-buy-researcher-silence-violate-labor-laws-critics-say.html> (accessed: 8th October 2021).
- Rogowsky, Marcus (2014) Analyse zur Sicherheit von Open-Source Software. Fakultät für Informatik, Technische Universität München. Retrieved from: https://www.net.in.tum.de/fileadmin/TUM/NET/NET-2014-08-1/NET-2014-08-1_20.pdf (accessed: 2nd September 2021).
- Rosenzweig, Paul (2012) Cybersecurity and Public Goods: The Public/Private "Partnership". Hoover Institution, Stanford University. Retrieved from: <https://bit.ly/3AxVjD5> (accessed: 10th August 2021).
- Samuelson, Paul A. (1954) The Pure Theory of Public Expenditures, *Review of Economics and Statistics* 36 (4): 387-389.
- Schulze, Matthias & Thomas Reinhold (2018) Wannacry About the Tragedy of the Commons? Game-Theory and the Failure of Global Vulnerability Disclosure, in: Audun Jøsang (Hg.), *Proceedings of the 17th European Conference on Cyberwarfare and Security*, Reading 2018.
- Schulze, Matthias (2019) Governance von 0-Day-Schwachstellen in der deutschen Cyber-Sicherheitspolitik, SWP-Studie, 10/2019, Berlin: Stiftung Wissenschaft und Politik.
- Siebert, Zora (2021) Digital Sovereignty - The EU in a Contest for Influence and Leadership, Berlin: Heinrich Böll Stiftung, <https://www.boell.de/en/2021/02/10/digital-sovereignty-eu-contest-influence-and-leadership> (accessed: 29th August 2021).
- Silomon, Jantje (2020) Bug Bounties: Bottom-up Initiatives as Forms of Cyber Arms Control? In: Brian K. Payne, Hongyi Wu (eds.) *Proceedings of the 15th International Conference on Cyber Warfare and Security*, Old Dominion University, Norfolk, Virginia, 12-13 March 2020, 431-438. Reading: Academic Conferences and Publishing International. DOI: 10.34190/ICCWS.20.137.
- Taddeo, Mariarosaria (2019) Is Cybersecurity a Public Good? *Minds and Machines* 29: 349-354.
- Thorne, Devin & Samantha Hoffman (31/8/2021) China's vulnerability disclosure regulations put state security first. Retrieved from: <https://www.aspistrategist.org.au/chinas-vulnerability-disclosure-regulations-put-state-security-first/> (accessed: 6th October 2021).
- Udemans, Chris (22/11/2019) China Working on Rules to Regulate Vulnerability Disclosures. Retrieved from: <https://technode.com/2019/11/22/china-vulnerability-disclosures-risks/> (accessed: 28th August 2021).
- Weber, Steven (2017) Coercion in Cybersecurity: What Public Health Models Reveal, *Journal of Cybersecurity* 3(3), 173-183.

Whittacker, Zack (05/11/2019) Huawei Calls Hackers to Munich for Secret Bug Bounty Meeting. Retrieved from: <https://techcrunch.com/2019/11/05/huawei-secret-bug-bounty-meeting/> (accessed: 28th August 2021).

YesWeHack (14/02/2019) YesWeHack Raises €4 Million and Plans to Disrupt Europe's Cybersecurity Market. Retrieved from: <https://blog.yeswehack.com/yeswehack-news/yeswehack-raises-e4-million-and-plans-to-disrupt-europes-cybersecurity-market/> (accessed: 28th August 2021).

YesWeHack (22/07/2021) YesWeHack Raises €16 Million to Accelerate its International Expansion. Retrieved from: <https://blog.yeswehack.com/yeswehack-news/yeswehack-raises-e16-million-to-accelerate-its-international-expansion/> (accessed: 28th August 2021).

Emerging Cyber risk Challenges in Maritime Transportation

Jussi Simola and Jouni Pöyhönen

University of Jyväskylä, Finland

jussi.hm.simola@jyu.fi

jouni.a.poyhonen@jyu.fi

Abstract: Maritime security and surveillance have become one of the main areas in managing overall situational awareness. For example, the growing importance of maritime traffic in cross-border trade has created new pressures to develop new technologies for accident prevention, especially in the ports. Maritime safety is also a matter of concern for continuity management. Automatic ship alarm systems, coastal radars and coastal cameras are not alone sufficient equipment to build maritime awareness. The Universal Shipborne Automatic Identification System (AIS) is a ship transponder system that is a globally used tracking system, but highly vulnerable to hacking. A major maritime traffic problem arises if transponders are switched off. Hybrid threats need coordinated hybrid responses; therefore, a cyber situational picture is also needed. Cyber situational awareness is an essential part of the management of maritime situational awareness. The lack of using real-time data from the maritime actors affects the correct formation of the common situational picture—for example, from the site of an accident. Cyber security is an essential factor in developing fairway navigation and all terminal (port-to-port) activities. This research will be done as a part of the SMARTER (Smart Terminals) -project that belongs to the SEA4VALUE program. The project aims to develop unique digitalized concepts that enhance safe transportation and reduce emissions in the port and the terminal areas. By using the multiagent system with sensor technology, e.g., in the harbors, it is possible to gather and share meaningful maritime security-related data. The study's primary purpose is to describe the operating environment and make an initial analysis of system requirements for optimizing situational awareness in the area of western ports of Finland.

Keywords: Situational awareness, port systems, cybersecurity, risks, information sharing

1. Introduction

“Maritime transport is a crucial activity for the European Union economy. It enables import and exports of goods, supply in energy, trade within the European Union and transport of passengers and vehicles. This activity relies on more than 1 200 seaports within the European Union, each with different organization, interests, challenges and activities.” (ENISA, 2019)

International and national maritime transportation systems are essential parts of critical global infrastructures. Digitalization and increased levels of autonomy in logistic transport chains are expected to take leaps forward in the coming years. This development can help create safer, more efficient, sustainable, and reliable service chains to meet the requirements for a better quality of life and global prosperity. Harbor operations connect the maritime transport to other modes of transportation and enable multimodal transportation. Smart harbors are in a central role in future transport logistics and supply chains. Well-build digital harbor infrastructure is essential in optimizing operations and planning for future investment and maintenance needs. Progressive data management and data sharing are the key issues for transparent, interoperable, safe, effective and environmentally friendly operations.

In Finland Smart Terminals (SMARTER) research project enlarges the scope of DIMECC Sea4Value program (DIMECC, 2020 a) to harbors and ports by develop solutions that benefit maritime transportation harbors in reducing emissions by optimizing harbor operations and improving cargo and people flow while improving the experience for all stakeholders. The mission of SMARTER is to create replicable models for digitalization, service innovation and data usage and sharing in the harbor environment and prepare for the future by taking steps towards smart and autonomous maritime transportation. (DIMECC, 2020 b).

At the same time of the development of digitalization, it also includes the existing maritime cyber environments. Therefore, it is necessary to address the relevant safety aspects of the maritime autonomy solutions. In any cyber environment, it is crucial that there are trustable information networks. In addition, the usability, reliability, and integrity of systems data needs to be high within the operating environment, where cyber security risks are continuously being highlighted by the threatening scenarios posed by the digital world. A modern society depends entirely on a cyber environment that provides dynamic services.

Especially the development of Information and Communication Technology (ICT) or Information Technology (IT) and Industrial Control System (ICS) or Information Technology (IT) systems will make different kinds of autonomy

solutions possible in all kinds of infrastructures. For example, in maritime systems the Universal Shipborne Automatic Identification System (AIS) is a ship transponder system that is currently used by most actors in the commercial shipping industry. The transponder transmits and receives information on VHF channels and thus enhance situation awareness in ships, but this globally used tracking system is highly vulnerable to hacking. For the maritime system of systems point of view, a major maritime traffic problem arises if transponders are forced to switch off in case of wrong operating. System-level threats are needed to be coordinated like hybrid responses; therefore, a system of systems-level cyber situational picture is also needed.

Digital assets of ports have two main digital assets groups. Information Technology (IT) systems, focused on the use of data as information, and Operational Technology (OT) systems, which use data to control and/or monitor physical processes (encompasses Industrial Control Systems including Supervisory Control and Data Acquisition and Distributed Control Systems) (ENISA 2019).

The SMARTER project aims are to develop unique digitalized concepts that enhance safe and effective transportation logistic in port and terminal areas. By using the multiagent system with sensor technology, e.g., in the harbors, it is possible to gather and share meaningful maritime security-related data. In that sense securing the cyber aspects of an interconnected system of maritime hosted by multiple stakeholders requires a system-of-systems view in cyber security. The main cyber security research question in this project is the following: How can a comprehensive cyber security architecture for smart port terminal be developed? However, we must first find an answer to the next question: Where can we direct the cyber security research to cover whole system of systems entity in port operation? A situation awareness from whole process elements is needed. Authorities and those involved in transporting goods and people need autonomous solutions and a logistical situational picture of both the port and the fairway. It is also about risk management, emissions, efficiency and continuity management. Consumers also need smooth port operations but a reliable picture of cargo and passenger traffic.

This paper provides a research approach for the investigation of cyber security at the system level and gives an answer to the question of where to direct cyber security research in the case of the SMARTER program. The paper emphasizes the importation of system description of the port process and its elements at the beginning of the research program. After that, the research will be able to continue to specify the comprehensive cyber security aspect of architecture like threat and vulnerability investigations, risks assessment and cyber security measures for the SMARTER project by the end of 2023.

2. The research objectives and port operations

2.1 Essential maritime security actors in the European Union

European Maritime Safety Agency (EMSA) works under the European Union and serves the EU maritime interests for a safe, secure, green and competitive maritime sector in Europe and worldwide. EMSA is the essential stakeholder in the maritime cluster in Europe and beyond. It provides services to the EU Member States and the Commission but also an innovative knowledge hub for the European maritime environment (European Maritime Security Agency 2022)

European Union Agency for Cybersecurity (ENISA) is the agency that works under the European Union purposed to achieve a high common level of Cybersecurity across Europe. ENISA contributes to EU cyber policy, aims to enhance the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperate intensively with the Member States and the EU bodies, and help Europe to prepare for the cyber challenges.

The International Maritime Organization (IMO) is the agency of the United Nations, which is responsible for measures to improve the safety and security of global shipping and prevent pollution from ships involved in legal matters, including liability and compensation issues and the facilitation of international maritime traffic (IMO 2019).

The International Maritime Cyber Centre of Excellence (IMCCE) consists of a Maritime Cyber Emergency Response Team (MCERT) and a cyber academy. The MCERT is an international cyber intelligence and incident support platform for the entire maritime ecosystem by providing international intelligence feeds, advice and support, including real-time assistance to members on cyber-attacks and incidents, and a Cyber Security

Reporting Portal (CSRP) for its members. The IMCCE will provide a focal point for the industry to help drive cyber awareness and response to cyber incidents (Container-News 2018).

2.2 Earlier investigations related to challenges of maritime security

Operators as Vessel Traffic Systems Operator tracks vessels to ensure that traffic is safe and safe distance between vessels is standard. Challenges of border guard operational work have listed below. Following challenges of formation of situational awareness have been investigated earlier by Simola & Rajamäki (2018). The challenges may be reflected more broadly in the port and fairways activities if physical and hybrid threats are not recognized.

1. The West Finland Coast Guard District is responsible for security in the whole sea area in Western Finland. The area of operations of the West Finland Coast Guard District covers the emergency area of four emergency centers.
2. The monitored area is quite broad, and they have to share the groups between the different workstations when a major accident occurs.
3. System-based obstacles in cross-border cooperation. Small ships or boats that are attempting to cross the Schengen border create additional challenges.
4. The situational picture of an individual patroller is based on the Virve communications and background information that has been collected before via radio communication.
5. Lack of automated functions and resources

Operational fieldwork covers statutory tasks involving the leading positions of other authorities as provided by the law on the Border Guard (Border Guard Act), such as executive assistance tasks and the management of Maritime Rescue.

2.3 Smart Terminals (SMARTER) research project in Finland

In Finland, the Sea4Value / Fairway (S4VF) research program has been established to create autonomy in maritime solutions (DIMECC, 2020 a). The first stage of the program concerns on automated remote pilotage fairway navigation. Now in the second phase of the program will leverage to harbor operations. This Smart Terminals (SMARTER) research project be enlarged and complemented to develop harbors and ports so that they meet the forthcoming needs of autonomous traffic and business. SMARTER has two main objectives. (DIMECC, 2020 b)

- The first is the reduction of emissions by optimizing port logistics.
- The second objective is to enable exceptional flow and experience for the passengers and cargo.

The structure of the project has been planned to have three use cases (Figure 1). The use cases are: Ship turnaround, truck traffic and passenger flow. Use cases are designed to support one another and there is a linkage between the use cases. The applied research work is organized into five work packages led by a responsible team of researchers. Cyber security research actions are included to Work Package 4 (WP4).

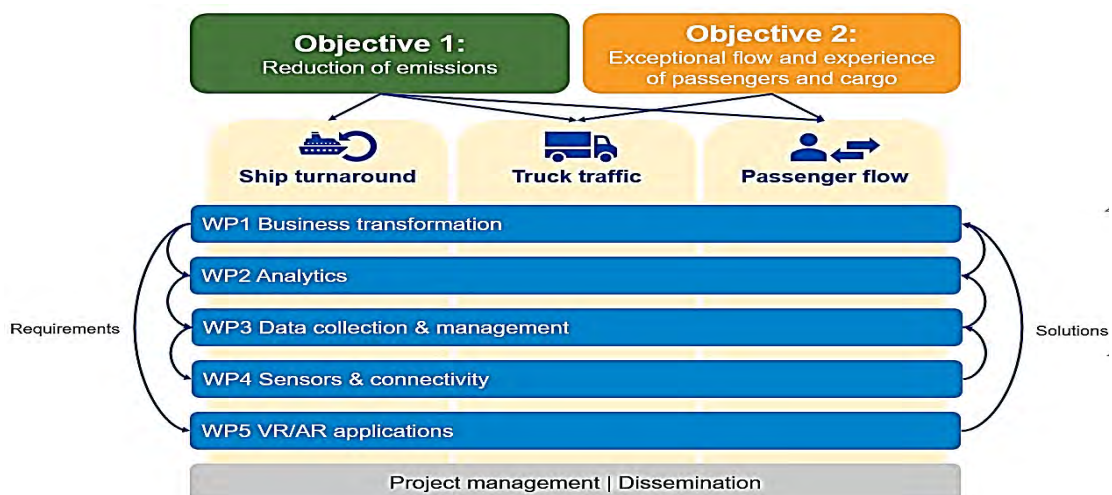


Figure 1: Project structure

In maritime transportations, the sea areas can be distributed to different approach areas of ports and ports itself. A ship proceeding to any port from the open sea through a fairway faces different demands during its transit, such as shallow waters, narrow shipping lanes, environmental effects and heavy traffic. After that a ship is ready to berth to a pier in the destination port and load or unload its goods, trucks/cars, or passengers. Port stakeholders and logistic systems are needed in successful port operations. Land transportation and its systems are included to port facilities. After that, it is also obvious that the regulatory requirements and needs for cooperations and communication with different stakeholders have been considered as part of port operations. In all cases of port processes, the information requirements and the amount of information needed are related to the accuracy and reliability of safety services. Figure 2 presents these processes. Cyber security awareness and information should cover all process elements.

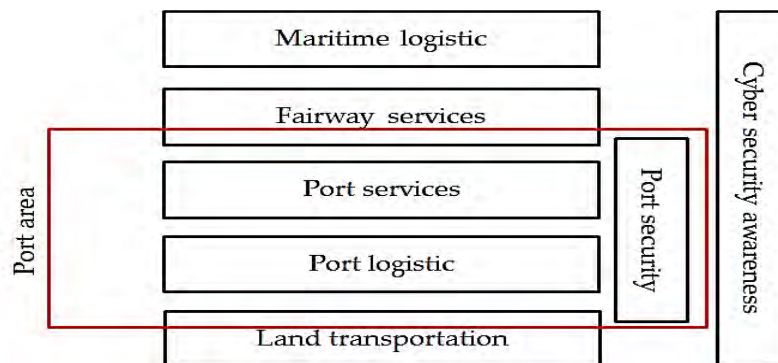


Figure 2: Port Processes

The research concentrates on examining and determinate impacting cyber risk factors and analyzing initial system requirements for the fairway navigation in the area of western ports of Finland. As figure 2. illustrates, the Port area consists of several functionalities. In addition to divided three sections, “Port Services, Port Logistic, and Port Security”, the port area expands its impacts to fairway services and land transportation. From the viewpoint of exploring the operating environment, it is necessary to understand the interaction between different functionalities. Working business continuity management means unexpected interruptions are removed or minimized. All sections of the process have to work under the umbrella of cyber security awareness. Cyber security is one crucial part of business continuity management.

The IMO Resolution Maritime Safety Community 428(98) states guidelines for Maritime Cyber Risk Management in Safety Management systems. Cybertechnologies are crucial to the operation and management of several systems critical to the safety and security of shipping and protection of the marine environment. The vulnerabilities of accessing, interconnecting, and networking of these systems may lead to realized cyber threats. (The, Maritime Safety Committee 2017).

A Mentioned above, a stable controlling system for transportation and other processes require a command and control or situation center that combines and analyze relevant data from the port, port facilities and functions. All kinds of practical information are needed. Monitoring different kind of threats is one crucial element, but it is required to monitor systems health also. Cyber domains have inner and outer vulnerabilities and threats. The port area consists of several of systems that have to be connected or integrated with each other. Efficient information sharing and updating situational awareness belong to each other. Improved situational awareness needs predictive functionalities that alert before any threats are realized.

3. Cyber security research framework

Traffic and transportation supply chains and their critical stakeholders’ systems are complex systems of systems characterized by a conglomeration of interconnected networks and dependencies. According to the EU Commission, ICT systems are significant parts of the operations and core processes. ICT systems are related to administration and to the management of information in the network. The components of the process levels also include ICS systems (EU Commission, 2009).

Martin C. Libicki (2007) has created a model of the cyber world based on The Open Systems Interconnection Reference Model (OSI), which has the following four layers: physical, syntactic, semantic, and pragmatic. Martti Lehto, a professor of cyber security at the University of Jyväskylä, has updated the four layers in Libicki’s (2007)

cyberworld model by adding a fifth layer that considers organizations' networks, ICT and ICS systems and the components of the systems (Lehto & Neittaanmäki, 2018). The five-layer structure for the cyber world of organizations is considered the research framework in this study on the cyber security of the port operations. An organization's cyber security operations require comprehensive awareness on the system level. The awareness of an organization and decision-makers can be seen as system-level awareness arrangement.

. It is possible to integrate an organization's three decision-making levels into a five-layer cyber structure in order to have a comprehensive system view of that organization's cyber security environment. It is a system-based approach to the topics and principles of an organizations comprehensive cyber security. The combination of system views, decision-making levels and an organization's cyber structure is described in Figure 3 (Pöyhönen & Lehto, 2020).

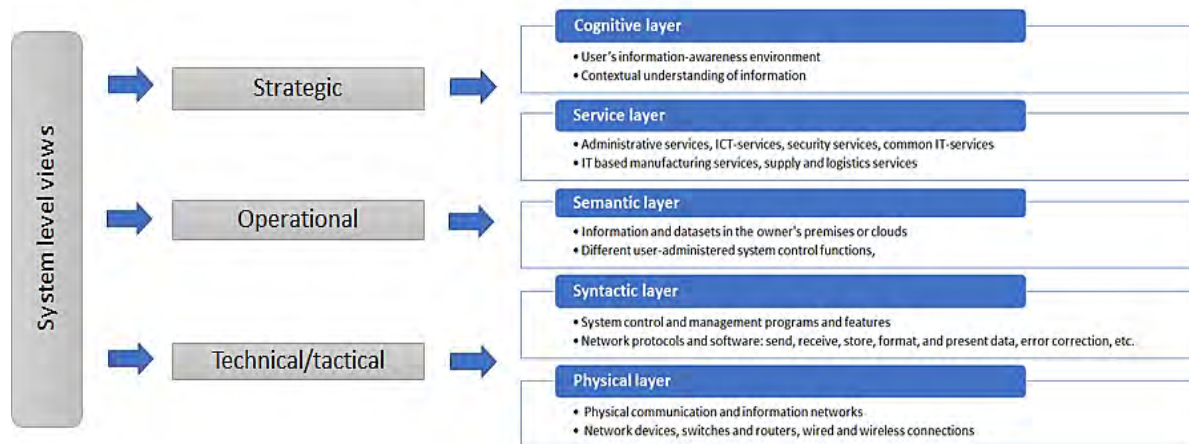


Figure 3: System-level view on organizational cyber security (Pöyhönen & Lehto 2020)

This case study is carried out with the guidance of Yin (2014). The case study illustrates the attempt to produce detailed information about the object being researched. The materials collected for this case study are based on publications, official reports, articles and other literary material.

The aim is to utilize the framework to define IT and OT ecosystem in the port environment that affects flowing port activities. Enhanced formation of situational awareness, information sharing and efficient decision-making processes are essential elements for the business continuity management. How and where to enhance and implement organizational cyber-physical security practices and how to create a hybrid secure umbrella for the port functionalities is the relevant issue in this initial research. As the simplified figure 3 illustrates, the maritime cyber-physical ecosystem consists of different systems and systems of systems. The crucial challenge encompasses several "system entities". Coherent formation of Maritime Situational Awareness requires functional interfaces from the systems. The separate system requirements of public authorities and commercial activities need to be coordinated. Vessel Traffic Services provider Finntraffic maintains maritime situational awareness, separate Control Center as Security Operations Center follow up their own situational picture and in addition to this, Coast Guard follow their own screens in the Maritime Rescue Command Center (Simola, Rajamäki 2018).

The following types of information and communication systems are used in maritime area for example; GOFREP (Gulf of Finland Reporting System), PORTNET Ship and port traffic notification system maintained by the authorities in Finland, Vessel Monitoring Systems (VMS), RFID (Radio Frequency Identification), AIS (Automated Identification System), VIRVE- KEJO-, POKE-equipment for authorities, RVT-mepe, NERCS, CleanSeaNet (CSN) the Union maritime information and exchange system, GPS-systems, Radar systems. SARSURPIC is the Search and Rescue for European Union authorities.



Figure 3: Communication Relationships between the Blocks in the Port Operations

4. Port operations and cyber security awareness

In their paper, “The Maritime Security Management System: Perceptions of the International Shipping Community” (2007), Vinh V. Thai and Devinder Grewal from Australian Maritime College emphasize the importance of identifying key shore-based and near-shore activities associated with maritime security management. The paper presents the findings of a research project on the Maritime Security Management System (MSMS). The principles of the paper can also be used for cyber security analyses in maritime environment. Then the main tasks are related to identifying key cyber security features associated with maritime operations. (Thai & Grewal, 2007, Pöyhönen, Kovanen & Lehto, 2021).

A cyber security situation model captures information about potential cyber threats against a system of systems (SoS) including operational and technical systems, an enterprise, a region, or a critical infrastructure (CI) sector. A comprehensive cyber security needs wide scale of analysis of a system of systems (or sub-system) against a set of threat events. Using threat scenarios makes it possible to get a picture of potential threats. This could materialize and result of harmful consequences and eventually lead thought to risk analysis and cyber security measures. Potential uses of threat scenarios should cover three scopes or scales involving systems of systems: the mission or business function, the enterprise, and the sector (or sub-sector) or region. (Bodeu & McCollum, 2018)

In the port research environment and in the cyber security investigation, there are three main processes, Ship Turnaround, Truck Traffic, and Passenger flow. On the other hand there are processes, relevant stakeholders, within organizations, operational regions and the maritime CI sector. Security elements like people, processes and technology are the key capabilities in the cyber environment, but at the same time they include vulnerabilities. In addition of processes and relevant stakeholders, people, processes, and technology are the third part of cyber security investigation. All these parts should be covered in order to have good situational awareness to relevant threats, vulnerabilities, risks and cyber security measures.

As an answer to the question: Where can we direct the cyber security research to cover the whole system of systems entity in port operation? the following consideration take place:

4.1 People: Stakeholders

Different surveys state that even though organizations have made significant security improvements, they have not kept pace with today’s determined adversaries and, as a result, many rely on yesterday’s security practices to combat today’s threats. It is obvious that a new model of cyber security that is driven by knowledge of vulnerabilities threats, assets, potential attack impacts, and the motives and targets of potential adversaries is required and extends to the entire maritime ecosystem from fairway piloting to all port operations.

Securing the cyber aspects of an interconnected system of maritime hosted by multiple stakeholders requires a system-of-systems view in cyber security. Digital transformation in providing safe maritime environment requires holistic approaches to analyze the whole operating environment. Digital transformation promotes reassessing and redesign of existing maritime services. The organizations of port stakeholders should have strategic, operative and tactic/technic views to the cyber structure of their own operational environment.

The strategic level, choices are primarily related to organization social responsibility, organization reputation, and ensuring business continuity (Finnish Standards Association SFS, 2016). The leadership is expected to make concrete strategic choices to guide and support the execution of relevant cyber situation awareness throughout the organization (Pöyhönen & Lehto 2020). The chosen measures should be comprehensively communicated to the organization's interest groups (Stouffer K., Falco J. and Scarfone K., 2011).

The measures at the operational level promote the strategic goals and thus comprehensive measures will increase situational awareness for holistic cyber security management. It must be based on risk assessment and analyses of the measures based on the assessment. It is also important that the organization declares and communicates the policy with which the leadership commits to the measures required to develop cyber security management. The declaration of a policy that ensures cyber security and the development of related procedures must be integrated with the organization's general policies (Pöyhönen & Lehto 2020). The highest organizational level is responsible for creating a policy that defines acceptable risk levels and the measures used in the reduction of risks (Stouffer et al., 2011). The maintenance of situational awareness regarding the cyber environment of the organization's processes makes possible to monitor and react efficiently to risks that constitute a threat within the organization's operating environment.

The tactical organization level encompasses the services and technologies that comply the processes with the structure of the cyber world (Pöyhönen & Lehto 2020). Consistent and predictable results are achieved more efficiently when operations are handled and managed as interrelated processes that function as a coherent system (Finnish Standards Association SFS, 2016). Cyber security threats and risks set special requirements for these processes and need for situation awareness at this level in addition to other operational requirements. In the cyber environment, the target can be achieved by defining the processes to be protected, choosing process control mechanisms successfully, and by using expedient technological solutions and services to protect the processes (Stouffer et al., 2011).

4.2 Process: Ship Turnaround, Truck Traffic, and Passenger flow

Processes are key to the implementation of an effective cyber security strategy. Processes are crucial in defining how the organization's activities, roles and documentation are used to mitigate the risks to the organization's information. Processes also need to be continually reviewed: cyber threats change quickly, and processes need to adapt with them. But processes are nothing if people don't follow them correctly (Dutton 2017).

4.3 Technology: Ship systems, Port systems, Port logistic systems, Truck Traffic systems

Technology is obviously crucial when it comes to cyber security. By identifying the cyber risks that your organization faces you can then start to look at what controls to put in place, and what technologies you'll need to do this. Technology can be deployed to prevent or reduce the impact of cyber risks, depending on your risk assessment and what you deem an acceptable level of risks (Dutton 2017).

ISO/IEC 27001:2013 (ISO27001) is the international standard for information security that specifies an information security management system (ISMS). The ISMS assists organizations manage their information security by addressing the people, processes, and technology (International Organization for Standardization, (ISO) 2013). NIST produces a voluntary set of guidelines for organizations against cybersecurity risks. NIST Cybersecurity Frameworks (CSF) does not replace ISO security standards, but it supports all other essential guidelines and guidances. The five simple core functions of the framework are as follows a) Identify, b) Protect, c) Detect, d) Respond, e) Recover (NIST 2018). It is usable for the different kinds of maritime stakeholders. The crucial challenge is that actors and their procedures to tackle cyber threats vary significantly in the shipping environment. Others have invested a lot of resources to protect systems, working methods and procedures at strategic, operational, and tactical levels. Others have done almost nothing to protect their working environment.

Effective working environments require effective cyber threat prevention mechanisms, but in addition to this, physical threats must have to ability to eliminate. Hybrid threats form a combination of these.

5. Discussion

In the SMARTER research project, the main aim is to have future ports more efficient than they have been before. That will happen to develop the key areas of ship turnaround, truck traffic and passenger flow. The development of these three main port processes includes the collection, production, and processing of cyber situation

awareness information related to all services that are needed in port operations. Cyber security is an essential part of these services.

Firstly, this paper presents a research framework for the study of cyber security in a maritime smart port operations. It is a five-layer cyber structure of an organization (Figure 3) that includes people, services, data, and ICT and ICS systems of the SMARTER process. Other components of it are the system views decision-making levels from every stakeholder's organization, including strategic, operational and tactical/technical levels. The five-level cyber structure and decision-making levels are needed to obtain a comprehensive system view of the cyber security situation from the cyber environment of SMARTER. That makes it possible to provide an answer to the research question: Where can we direct the cyber security research to cover the whole system of systems entity in port operation? It is needed the system thinking approach to organization five layer cyber structure. It makes it possible to have holistic cyber security architecture using the research framework mentioned earlier in this paper.

6. Conclusion

This paper provides the first research approach for investigating cyber security at the system level in a smart port process of autonomous activities development of maritime operations. It emphasizes the importation of cyber security awareness at all system systems levels at the beginning of the research. The paper emphasizes system thinking using a presented cyber security research framework.

Security elements like people, processes, and technology are the key capabilities in the cyber environment, but at the same time, these capabilities also include vulnerabilities. Vulnerabilities and threats can be divided into internal and external sources. Situation information from port processes, relevant stakeholders, and technologies behind systems, like port systems, port logistic systems and traffic systems, are needed to cover risk assessment of all port environments. Monitoring IT systems and sub-systems health is an essential part of this procedure. Therefore workable situation center or SOC that handles cyber-physical threats from the maritime ecosystem is needed in the port environment. Artificial Intelligence-based solutions will replace human capabilities shortly. Technological development will challenge traditional monitoring systems. It is not enough that humans follow the flow of information from simple screens; cyber situational awareness functionalities must ensure business continuity.

There is also a need to gather and analyze data from the environment and potential threats from outside the port. Cooperation between existing situation centers has to enhance and rationalized. The Finntraffic (VTS operator) and The Maritime Rescue Coordination Centre (MRCC) have a crucial information gathering, sharing and operational role in this environment, not forgetting Custom, Police and rescue services. Efficient threat information sharing is a crucial part of situational awareness that ensures business continuity management. It is also essential that shared information is understood in the same way between different actors.

There is a need for standardized models for digitalized systems and processes, procedures, data handling etc. In the near future, automated artificial intelligence solutions with multiagent systems constitute an opportunity but also threats. The risks of false alerts by AI surveillance systems must be minimized when programming new solutions for the port area. Efficient strategical, operational and tactical situational understanding needs maritime computer emergency response functionalities that cover all system enterprises that are in use. Threats and vulnerabilities have to tackle if the purpose is to reduce pollution and enhance the flow of passengers and goods traffic.

In Finland, Official operators as Vessel Traffic Services have to follow, e.g., the Vessel Traffic Service Act (The Finnish Parliament 2005). The act obligates to inform authorities in all abnormal threat situations. Before organizational changes and artificial automation, Port Facilities need to organize precise and common guidelines on preparing and preventing cyber and physical threats. The combination of system of systems has to control with the same principles. This does not necessarily mean unifying systems into a single system. It has to be possible to disconnect the infected functionalities from the system chain and the system recovery must be capable. Optimizing harbor operations and improving cargo and people flow are essential purposes of the SMARTER project. System-level thinking cannot be separated from each other level. Strategical and operational level understanding about the ecosystem has to support tactical level purposes and vice versa. Horizontal and vertical situational awareness must be at the same level of understanding.

All these parts should be covered in order to have comprehensive situation awareness to relevant threats, vulnerabilities, risks and cyber security measures. After this, the work will continue towards answering the main research question of the SMART project: How can a comprehensive cyber security architecture for smart port terminal be developed?

References

- Bodeu, D. J. & McCollum, C. D. (2018) System-of-systems threat model. The Homeland Security Systems Engineering and Development Institute (HSSEDI) MITRE: Bedford, MA, USA
- DIMECC Oy, (2020a) SEA FOR VALUE (S4V). 12.2.2020. Available on 3 December 2021: <https://www.dimecc.com/dimecc-services/s4v/>
- DIMECC Oy (2020b) Project proposal for One Sea – autonomous maritime ecosystem DIMECC Sea4Value Smart Terminals (SMARTER)
- Dutton, J. (2017) *Three pillars of cyber security*. -09-26T10:00:51+00:00, [viewed Jan 12, 2022]. Available from: <https://www.itgovernance.co.uk/blog/three-pillars-of-cyber-security>.
- European Maritime Security Agency (2022) *This is EMSA*. Available from: <http://www.emsa.europa.eu/about.html>.
- European Union Agency for Cybersecurity ENISA (2019) PORT CYBERSECURITY. Good practices for cybersecurity in the maritime sector. NOVEMBER 2019
- Finnish Parliament (2005) Vessel Traffic Service Act 623/2005. Ministry of Transport and Communications
- Finnish Standards Association SFS (2016) Johdanto laadunhallinnan ISO 9000 -standardeihin. Available on 3 December 2021: slideplayer.fi/slide/11133323/
- The first International Maritime Cyber Centre of Excellence (2018) -10-17T09:00:20+00:00 [viewed Jan 12, 2022]. Available from: <https://container-news.com/iinternational-maritime-cyber-centre-excellence/>
- International Organization for Standardization, (ISO) (2013) *ISO/IEC 27002:2013 Security techniques — Code of practice for information security controls*. ISO. Available from: <https://www.iso.org/standard/54533.html>.
- Lehto, M. & Neittaanmäki, P. (2018) *The modern strategies in the cyber warfare*. Cyber Security: Cyber power and technology. Berlin: Springer.
- Libicki, M. C. (2007) *Conquest in Cyberspace – National Security and Information Warfare*, Cambridge University Press, New York 2007.
- Maritime Safety Committee (2017) *Maritime Cyber Risk Management in Safety Management System Resolution MSC.428(98)*, 16 June.
- NIST (2018) *The Five Functions*. -04-12T13:27-04:00, [viewed Jan 7, 2022]. Available from: <https://www.nist.gov/cyberframework/online-learning/five-functions>
- Pöyhönen J., Kovanen T. & Lehto M. (2021) Basic Elements of Cyber Security for an Automated Remote Piloting Fairway System. The proceedings of the 16th International Conference on Cyber Warfare and Security ICCWS2021, 25 - 26 February 2021, A Virtual Conference, Tennessee Tech University and the Oak Ridge National Laboratory, USA, pages 299-308
- Pöyhönen, J. & Lehto, M. (2020) Cyber security: Trust based architecture in the management of an organization security. The proceedings of the 18th European Conference on Cyber Warfare and Security ECCWS2020, 25-26 June 2020, University of Chester, UK, pages 304-313
- Simola, J. and J. Rajamäki (2018) Improving Cyber Situational Awareness in Maritime Surveillance. JOSANG, A., ed. *the 17th European Conference on Cyber Warfare and Security ECCWS 2018*. Oslo, Norway,.
- Stouffer K., Falco J., Scarfone K. (2011) NIST Special Publication 800-82. Guide to Industrial Control Systems (ICS) Security. Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce. Available on 3 December 2021: <http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf>
- Thai, V.V. & Grewal, D. (2007) *The Maritime Security Management System: Perceptions of the International Shipping Community*, 2007. Article in *Maritime Economics & Logistics* · June 2007.
- Yin, R.K. (2014) *Case Study Research, Design and Methods*. 5th ed. Thousand Oaks: Sage Publications.

Improving Protection Against Cybersecurity Attacks of Emergency Dispatch Centers

James Sweeney and Vu Tran

Capella University, MN, USA

Jsweeney1@capellauniversity.edu

vu.tran@capella.edu

Abstract: Public service answering points (PSAPs), also known as 911 dispatch centers, serve as integral and critical infrastructure components that serve as a conduit between the public and emergency assistance such as police, fire and emergency medical services. It has been demonstrated in previous research, that PSAPs may be exceptionally vulnerable to telephonic and distributed denial of service (TDoS/DDoS) attacks with potentially devastating effects. It is the purpose of this study is to gather the best practices from experts and tap into the knowledge of the professionals tasked with safeguarding PSAPs every day. Due to the unknown full capabilities of the PSAPs and the antiquated infrastructure on which they must operate, it is unclear just what safeguards are in place to defend these critical infrastructure components against attacks of this nature. To gather this information, a multiple rounds qualitative Delphi study was conducted. Through this process, participants were asked to comment on current tactics and techniques, practices that could be implemented under ideal conditions without political or financial hurdles and how to bridge the gap between current and optimal environments. PSAP administrators, hereby referred to as experts, with a minimum of five years of experience working within a United States PSAP were included in this study and provided a firm understanding of their capabilities. The suggestions provided by participants included patch managements, updated hardware, federally mandated standards, regular plan exercises, and standardized education. After gathering and analysing the data, three basic tenets could be appreciated including cyber hygiene, preparedness and intelligence, and education and training. It is expected that the results of this study will prove integral in not only better securing the PSAPs within the United States critical infrastructure, but also understanding some of the hurdles and difficulties PSAP administrators must overcome.

Keywords: 911 Dispatcher Center, DDoS, TDoS, 911 communications center, PSAP, Public Service Answering Point

1. Introduction

Public service answering points (PSAPs), also known as 911 dispatch centers, serve as the gateway for those needing to access emergency assistance from police, fire, and emergency medical services (FCC, 2021). The delivery of emergency services through the 911 system is crucial for ensuring health, safety, and welfare of the public. The proper emergency services are dispatched after the caller communicates with a telecommunications officer at their local PSAP. In the United States, users can dial 911 and be immediately put in contact with telecommunicators (i.e., dispatchers) who provide pre-arrival instructions and allow early treatment and intervention be administered before the arrival of emergency units (FCC, 2021). According to the National Emergency Number Association (NENA), an estimated 240 million 911 calls are placed in the United States every year, with approximately 80% being placed via wireless devices (NENA, 2021). Calls for assistance are directed through primary and secondary PSAPs equipped with personnel trained to provide emergency assistance and serve as a conduit between the caller and the field unit (FCC, 2021).

As technology advances, PSAPs are beginning to transition to the Next Generation 911 (NG911) systems, which are Internet Protocol (IP) based. This allows for the transmission of text messages and Voice-over-IP (VoIP) calls to be seamless through the 911 network (911.gov, 2016). This allows a caller to summon help when they might not otherwise have been able, for instance when being followed and cannot call without putting themselves at risk. Even though this added functionality lowers the threshold for access, it introduces a significant vulnerability for cyber-attack (Kang, Gilgor, & Sekar, 2017). The upgrading of this system is estimated to cost over \$10 billion and anticipated to take multiple years (Jackson, 2017). The cybersecurity challenges to PSAPs are here today, yet the realization of the NG911 is years away. Due to the critical nature of the 911 service, a cybersecurity solution to bridge the gap between today's system and NG911 is needed (APCO International, 2017).

The purpose of this qualitative Delphi study is to gather the best information security practices of experienced professionals safeguarding PSAPs daily. From that data, a model can be developed to be used by PSAPs nationwide until NG911 become available. Today's PSAPs are based on an antiquated architecture that has been in place since the 1970s (Jackson, 2017). The lessons learned from this intermediate step will provide valuable information and help foreshadow potential issues to avoid during the implementation of the new infrastructure. This model can be used as a vehicle to finetune the viability of the long-term solution. Our research question:

What model can be developed to better safeguard PSAPs against cyber-attacks while they still function using the current antiquated architecture?

2. Background

According to the NENA, an average of 650,000 emergency calls for assistance are placed through PSAPs daily (NENA, 2021). Calls are placed via landline or cellular device through their respective infrastructures (Guri, Mirsky, & Elovici, 2016). The 911 network serves as an entry point and conduit for the public to access the emergency services system. It has become a vital component to emergency response and disaster preparedness, functioning as a backbone that enables field units to receive and convey crucial information (FCC, 2021). Public service answering points have been recognized as a component of the critical infrastructure sectors through the Uniting and Strengthening of America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act (USA Patriot Act; USA Patriot Act, 2001). Furthermore, the communications sector is unique, as its failure has the potential to affect multiple other sectors (White House, 2013).

The current infrastructure on which PSAPs rely was initially developed in the 1970s and is antiquated compared to current day standards and must be updated to get on par with what everyday citizens can do with a smartphone (Jackson, 2017). Researchers from Ben-Gurion University of the Negev Cybersecurity Research Center (Guri et al., 2016) conducted a study to determine how vulnerable a PSAP is to a telephonic denial of service (TDoS) attack, and the devastating effects should an attack be successful. A botnet consisting of approximately 6,000 bots could launch a TDoS attack paralyzing the PSAPs in a state the size of North Carolina for days (Guri et al., 2016). Furthermore, a botnet consisting of approximately 200,000 bots using the same attack would have the ability to affect the 911 system throughout the entire country. As the abilities and functionalities of botnet variants continue to grow, this possibility becomes increasingly real. As of 2017, over 40% of PSAPs and critical infrastructure sectors have experienced some level of cyberattack (Seals, 2017). The federal government is taking steps to develop tools that may be implemented to defend these assets against such attacks. It is reported that one out of 300 emails to public administration contain malicious data with 184 cyber-attacks made on the public sector from 2016-2018 (Motorola, 2019).

The money needed to upgrade the system is substantial and therefore prohibitive. The Association of Public-Safety Communication Officers (APCO) reports a substantial grant program from Congress and many years are needed to accomplish this initial upgrade (Jackson, 2017; Kang et al. 2017). It is estimated at least \$10 billion is needed to provide all PSAPs nationwide with the resources needed to enhance their systems and ensure proper mechanisms are implemented (Jackson, 2017). This situation requires that a solution be developed and implemented to bridge today's infrastructure to the proposed NG911 infrastructure. To facilitate this, the Cybersecurity and Infrastructure Security Agency (CISA), in conjunction with the SAFECOM-NCSWIC Next Generation 911 working Group, relies on stakeholder feedback to develop and innovate new concepts to safeguard the 911 Communications System (CISA.gov, 2021). CISA has also developed several resources highlighting the vulnerability of the 911 system and how to enhance the overall cybersecurity (CISA-Transition, 2021).

A challenge we face in search for a bridge gap solution for PSAPs problem is the lack of uniformity in practices and techniques currently used in PSAPs. As the result, we decided to focus this study on understanding best practices at these PSAPs.

3. Methodology

A qualitative Delphi method was utilized for our study. This method relies on an iterative interview process to develop a consensus among the participants of the study (Skulmoski, Hartman, & Krahn, 2007). Multiple rounds of interviews were conducted, with feedback interspersed between each round. The process completed when consensus was reached among participants (Skulmoski, Hartman, & Krahn, 2007). If there were disagreements that could not be resolved, the disagreements were recorded. With each new round, new questions were developed based on the responses from the previous round and presented to participants to build consensus. Our role as researchers is to facilitate the consensus building process. During the interview process, we focus on clarifying the interview process, asking the interview questions, collecting responses, and disseminating results for participants' approvals. We point out the disagreements among participants to help facilitate a resolution. All communications are strictly between individual participants and the facilitators. Participants do

not directly engage each other during the interview. Convergence was reached when there was a final agreement on a list of security measures for PSAPs or when no further consensus can be reached.

According to Hasson, Keeney, & McKenna (2000), study's validity using a qualitative Delphi method is assured by several mechanisms. The method relies on a collaborative decision-making process involving multiple participants which is more likely to yield correct outcomes than a single person. Decisions are subject to multiple rounds of reviews and challenges to further enhance their validity. The use of subject-matter experts as participants strengthens decision validity. Keeping participants from directly engaging each other during the rounds of interview helps prevent the risk of early decision convergence due to peer pressure.

The Delphi methodology has been used to develop models within the cybersecurity space before. Karabacak, Yildirim, and Baykal (2016) developed a model that could be used to measure the readiness of national critical-infrastructure protection efforts based on a grounded theory evaluation of national cybersecurity projects. Turoff, Bañuls, Plotnick, Hilitz, and Ramírez de la Hueraga (2016) used the same methodology to develop a single model exhibiting the collective viewpoints regarding the interactions of emergency managers and critical infrastructure conditions. Seppänen, Luukkala, Zhang, Torkki, and Virrantaus (2018) queried participants through a Delphi study that examined the interdependencies between the failures of critical infrastructures as perceived by regional preparedness committees. Kong, Kim, and Lee (2017) conducted a Delphi study to create a model that can be used to classify cyber-attackers and evaluate their capabilities.

3.1 Participants

Below are the participation criteria:

- PSAP administrators, supervisors, and decision-makers responsible for deciding cybersecurity techniques and tactics best used securing their centers. These participants' daily function is to oversee the security and operation of the PSAP systems and have an intricate understanding of how each supporting system interacts.
- Administrators who have worked in their respective positions for at least five years.
- Administrators who have worked at PSAPs within the United States.
- Participants must currently hold a basic telecommunicator certification (BTC) or equivalent, with preference given to those possessing emergency medical dispatch certifications.
- Participants must be willing and able to voluntarily participate from the study's beginning until the end. Participants who withdraw from the study early will have their interview data removed from the study.

We looked for 15 PSAP professionals to participate. This number met the recommendation of between 10 to 50 participants (Iqbal & Pison-Young, 2009). The following steps were taken during the recruitment process:

- A presentation reviewing the study was provided to state-wide Emergency Medical Services 911 sub-committee meeting. The terms and conditions of permissions and ethical considerations were provided.
- PSAP administrators were found and contacted through LinkedIn. The same information from the initial presentation was conveyed to these potential participants.
- Potential participants who express interest were vetted by the researcher to ensure compliance with the requirements.
- All participants were contacted and notified whether they qualify. If qualified, additional information was provided regarding when the study was scheduled to be performed.

Before conducting the study or asking any questions, consent forms were obtained electronically.

3.2 Setting

A search on LinkedIn was completed to access participants nationwide. Names, locations, and agency affiliation of the participants were withheld for security purposes. The setting for the first three rounds of the interview was done using a combination of online survey, specifically SurveyMonkey® and emails. The fourth round was conducted via Zoom and phone interviews.

3.3 Analysis of Research Questions

Three main interview questions were posed to participants in Round One interviews:

- What cybersecurity standards and techniques (both internally and externally) have you implemented to secure your PSAP against cyber-attacks?

- In an ideal world in which finances, politics, and dated infrastructure were not issues, what other techniques and tactics would you like to see implemented to better safeguard your PSAP against cyber-attacks while still operating using the current network infrastructure?
- With the gap identified and working with the current infrastructure, what steps can be taken that would better protect PSAPs from potential cyber-attacks, and what would it take for you to implement these practices?

4. Data Collection and Analysis

The responses to Round One of the interviews were open-ended to allow participants freedom to explain and expand upon their responses. The open-ended responses received from SurveyMonkey were entered into NVivo 12 (NVivo, n.d.) for data coding and content analysis. Participant responses were reviewed and categorized according to the topic of the response. For example, one response recommended firewalls, strong passwords, and anti-virus protections. This response was coded once for firewalls, once for passwords management and another for anti-virus protection. Emerging patterns and themes were developed to guide the development of interview questions in subsequent rounds. Responses to these subsequent interview questions allowed the researcher to further refine the patterns and themes from the previous round into concrete security measures. Through the process of consensus building over several interview rounds, researchers developed the final list of security measures for enhancing existing PSAPs. Within each round of interviews, the participants were provided with the following:

- A report and explanation of the response analysis from the previous round.
- Items as uncovered in the previous round to be further evaluated in the next round.
- For each item identified, an aggregated score representing the collective agreement of participants agreeing with the item from the previous round.
- A list of questions for exploration and expectations of the new round of interview.

The participants provide responses to the interview questions in each round. For Round One, Two, and Three, written responses were provided. For Round Four, Zoom or phone calls were used.

5. Findings

5.1 Participants

Seventeen potential participants responded, with one participant not meeting the length of service requirements. Instead, that person passed the survey to someone else (P9) who did fit the requirements. P17 responded with interest but decided to not participate due to time constraints. Data collection began with 16 participants, which allowed for the loss of some participants should some decide not to participate. The rounds were continued until a minimum of 10 participants responded.

Table 1 Description of Potential Participants

Identifier	Role	Years	Within US	Certification
P1	PSAP Administrator	20+	YES	YES
P2	PSAP Administrator	5+	YES	YES
P3	Communication Supervisor	5+	YES	YES
P4	CAD Administrator	10+	YES	YES
P5	Communications Division Manager	10+	YES	YES
P6	Assistant Director	10+	YES	YES
P7	PSAP Administrator	30+	YES	YES
P8	System Analyst	5+	YES	YES
P9	Director	15+	YES	YES
P10	Executive Director	NO	YES	YES
P11	Emergency Communications Center Manager	10+	YES	YES
P12	911 Communications Technical Systems Coordinator	10+	YES	YES
P13	State 911 Program Manager	15+	YES	YES
P14	PSAP Administrator	10+	YES	YES
P15	PSAP Administrator	25+	YES	YES
P16	Director of Information Technology	10+	YES	YES
P17	Emergency Communications Specialist	5+	YES	YES

5.2 Round One Interview Results

Each participant provided the researchers a list of responses for each interview questions posted on SurveyMonkey®. The researchers collected and distributed all the responses to all participants via email. Participants were asked to review all the responses and categorize them into core themes. Researchers reviewed the themes and suggested how they can be organized. Except for their own responses, the participants do not know the owners of other responses. The responses to all three questions were categorized into the several core themes by the participants after several rounds of email exchanges (Table 2).

The final set of themes were presented via SurveyMonkey as a survey for participants to vote. Participants were asked to elaborate on each theme as to what security measures they previously implemented along the themes identified. Table 3 captures the participants' responses regarding specific security measures implemented in the context of the common themes identified from their responses to our Question 1. A total of 64 suggestions were gathered through participant interviews. Participants were also asked to categorize these security measures (Table 3, Column 2).

Table 2: Themes Generated by the Participants to Questions 1, 2, and 3 of Round One Interview

Item	Based on Question 1 Responses	Based on Question 2 Responses	Based on Question 3 Responses
1	Properly configured Firewalls	Implementation of defense in depth methodology such as use of redundant firewalls, independent intrusion detection monitoring systems.	A statewide security operations center to monitor all PSAP systems.
2	Maintaining multiple ISPs	Complete separation of network infrastructure at all levels.	Accepted best standard.
3	External devices prohibited in the towers	Dedicated IT person.	Federal mandate.
4	Network isolation	Nothing can be done on this current platform.	No idea / there is nothing that can be done.
5	Limited Internet access	ESINet Migration.	Education and training (including tabletop drills).
6	Control traffic flow	State-wide cybersecurity program to coordinate efforts and provide training.	Secure network (network isolation, disabling unused ports, etc.)
7	Limit Admin access	Additional network monitoring outside of firewall.	Strong vulnerability assessments.
8	Traffic monitoring via port and IP address	Better virus scanning.	Information sharing among PSAPs.
9	Redundancy	Public safety only DNS servers geographically separated.	Frequent good vulnerability assessments.
10	Use of a VPN to secure data	Stiffer laws and penalties.	Active monitoring.
11	Incident response plan development	Updated server OS.	Funding.
12	Strong password standards	Network monitoring.	External expertise to supplement internal resources and personnel.
13	Anti-virus protection	Implementation of defense in depth methodology such as use of redundant firewalls, independent intrusion detection monitoring systems.	System redundancy.
14	Internet of Things Awareness training	Complete separation of network infrastructure at all levels.	The best fire walls possible to protect the 911 network as we move into NG
15	Implementation of NIST Framework	Dedicated IT person.	Enhanced and upgraded equipment (Specifically firewalls).

Table 3: Themes identified from Question 1 (with Votes) by Participants

Response	Vote
Properly configured firewalls	3
Maintaining multiple ISPs	3
External devices prohibited	3
Network Isolation	3
Limited Internet access	2
Control traffic flow	2
Limit admin access	2
Traffic monitoring via port and IP address	2
Redundancy	1
Use of a VPN to secure data	1
Incident response plan development	1
Strong password standards	1
Anti-virus protection	1
Internet of Things Awareness training	1
Implementation of NIST Framework	1

Table 4: Security Measures in Responses to Question 1, with Category

Security Measure	Category
Administrative Internet access	Admin Access
Access to the systems is controlled via security log on	Admin Access
Anti-virus protections	Anti-Virus
No other items are permitted to be plugged into towers	External Devices
Do not allow flash drives or other input devices	External Devices
USB port restrictions	External Devices
Firewalls	Firewall
Firewalls on critical servers	Firewall
Firewall monitoring	Firewall
Incident response plans	Incident Planning
Internet access is disabled	Limited Internet
No external Internet access	Limited Internet
Internet of Things Awareness	Education
Systems are on their own networks	Isolation
Computers are closed off from Internet	Isolation
Network is isolated with implicit deny rules in the firewall	Isolation
Just started implementing some NIST	NIST
Rollover of incoming 911 calls	Redundancy
Security against TDoS attacks are handled through our 911 System service provider	Service Provider
Maintaining multiple ISP connections	Service Provider
Separate paths and different ISPs	Service Provider
Strong password standards	Strong Passwords
One-way connection	Traffic Flow
Only servers have outgoing access	Traffic Flow
Appliances at the edge that monitor traffic	Monitor
Active monitoring outside the firewall.	Monitor
Our IT department has installed VPN	VPN

This process for arriving at common themes was the same for Questions 2 and 3. The resulting security measures are summarized in Tables 4 and 5.

Table 5: Security Measures in Responses to Question 2, with Category

Security Measure	Category
Complete separation of network infrastructures	Network Isolation
Further segmentation of the administrative network	Network Isolation
Having an IT person who is a department employee	Personnel
A staffed NOC allowing technician to be notified	Personnel
Inhibit the external access to internal systems	Defense in Depth
Multiple redundant firewalls... fully automatic patch monitoring	Defense in Depth
Additional sensors on the network with more advanced IDS	Defense in Depth
Public safety only DNS server	DNS
This person would provide needed training	Education

Security Measure	Category
Preparing to migrate to an ESInet	ESInet
I would like network monitoring	Network Monitoring
Not sure there is a good solution for TDoS	Nothing
Implement a state-wide cybersecurity program	State-wide program
Better laws, stiffer penalties	Laws
Updated server operating systems	Updates
Better virus scanning	Defense in Depth

Table 6: Security Measures in Responses to Question 3, with Category

Security Measure	Category
It takes active monitoring	Monitoring
Education and training is vital	Education
Provide cyber security training	Education
Best firewalls possible to protect the 911 network	Firewall
State level funding	Funding
It is going to take funding	Funding
Federal mandate and funding	Funding
Funding and external expertise	Funding
Sharing real time information	Info Sharing
Standard that is backed by a federal mandate	Mandate
External expertise to supplement internal resources	Personnel
It takes active monitoring and knowledgeable people	Personnel
It is going to take funding and redundancy	Redundancy
It would take an accepted best standard	Standard
Statewide security operation center	Statewide center
Good vulnerability assessments	Assessments
Secure network, disabling unused workstation ports	Cyber Hygiene
Continual monitoring and upgrades/updates	Updates

5.3 Round Two Interview Results

During Round Two interview, participants were asked to confirm whether the suggestions gathered in Round One were acceptable. Of the 10 Round Two responses, 90% of the participants agreed the suggestions were acceptable. One participant stated they were not but did not offer any reasoning why or offer any additional suggestions. Once the list was accepted, some themes became increasingly obvious.

During the analysis, three overarching categories could be noticed. A vast majority of the suggestions were based on employing strong cyber hygiene practices like network monitoring, system redundancy, virus scanning, and password standards. A total of 41 of the 64 suggestions could be categorized as cyber hygiene practices, representing a percentage of 64.0625% of suggestions, shown in Table 6. Other suggestions spoke to the preparedness and intelligence of the PSAP and its ability to respond to an incident. This category included suggestions such as frequent and strong vulnerability assessments, state-wide coordination efforts, and information sharing among PSAPs. A total of 17 suggestions could be categorized under preparedness and intelligence, resulting in 26.56%, shown in Table 7. Finally, education/training was another theme that was appreciated through participant suggestions. Three suggestions spoke to standardized awareness training, representing 4.6875% of the total number of suggestions, shown in Table 8.

Table 7: Cyber Hygiene Category

Category	Number
Complete separation of network infrastructure at all levels.	7
Additional network monitoring outside of the firewall.	4
Properly configured firewalls	4
Limited Internet access	4
Multiple Internet Service Providers to increase bandwidth.	3
External devices prohibited in towers	3
Enhanced and upgraded equipment	3
Improved virus scanning.	2
System redundancy	2
Control of traffic flow	2
Implementation of defense in depth methodology	2
VPN Use to secure data	1

Category	Number
Strong password standards	1
Implementation of NIST framework	1
Geographically separated, public safety specific, DNS servers	1
Updated server operating systems	1
TOTAL	41
PERCENTAGE	64.0625%

Table 8: Preparedness & Intelligence

Category	Number
Funding	4
External participants to supplement internal resources and personnel	4
State-wide cybersecurity program to monitor and coordinate efforts	2
Frequent and strong vulnerability assessments.	2
Information sharing among PSAPs.	1
Stiffer laws and penalties for bad actors.	1
Federally mandated cybersecurity protocols.	1
Accepted best practice and standards.	1
Incident response plan development.	1
TOTAL	17
PERCENTAGE	26.56%

Table 9: Education & Training

Category	Number
Incident plan education, training and exercising	2
Standardized training	1
TOTAL	3
PERCENTAGE	4.6875%

5.4 Round Three and Round Four Interview Results

In Round Three and Four, participants were asked whether they agreed with these categories and if they would make any additional suggestions or changes. Out of all the responses, all participants agreed with the recommendation organization or content. Some participants failed to provide an answer to some of the questions. 10 participants remained through the entire interview process. The consensus reached among these participants was around 80% across all three interview questions provided.

Table 10: Agreements to the Round Two Results

Response	Question 1	Question 2	Question 3
Yes	8	8	7
No Answer	2	2	2
% In Agreement	80.00%	80.00%	77.78%

5.5 The Model

According to Turoff, et al., a model is a tool that outlines a process for implementing a strategy to solve a problem (Turoff, Bañuls, Plotnick, Hilitz, & Ramírez de la Hueraga, 2016). Based on the data gathered, the model proposed contains three areas of cybersecurity improvements including, Cyber hygiene, preparedness & intelligence, and education & training. It is surmised implementation of this model can assist PSAP administrators in securing centers against cybersecurity threats in the near term as they wait for the NG911 technology be implemented and infrastructure upgraded. The safety measures for each area of the model is listed in Tables 7, 8, and 9.

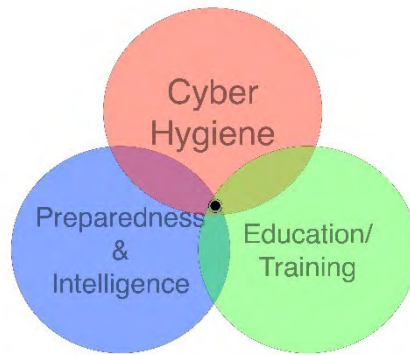


Figure 1: Cybersecurity Model for PSAPs

6. Discussion, Recommendations and Conclusions

6.1 Discussion

The purpose of this study was to tap into the expert knowledge of participants to gain a consensus on techniques and tactics that could be used to develop a model for addressing near term security needs of PSAPs. Our model identifies three safety measure areas: cyber hygiene, preparedness and intelligence, and education and training. Cyber hygiene includes common cybersecurity practices like defense in depth, strong password requirements, and network isolation. Preparedness and intelligence encourage PSAP administrators to share information and work collaboratively in defending against cyber-attacks. Finally, education and training provide a standard for all telecommunicators and management to better protect this infrastructure.

6.2 Recommendations

The model developed provides insights into the near-term security measures needed by today's PSAPs. This model can be used as a check list for analyzing the cybersecurity readiness of each operating PSAP center in the United States. The model can be used to help prioritize the implementation of security measures in the NG911 program rolled out in the future. The model can be used in reviewing historical data on PSAP attacks to identify potential gaps between what attacks performed and what security protections the model recommends.

This study should only be considered the starting point as further studies are needed to validate the model. Our study relies on the opinions and consensus of a panel of ten PSAP professionals. Additional interviews with PSAP professionals can help further confirm the model's list of security measures. Reviewing the list of security measures captured in the model against security standards such as NIST will allow identification of gaps between the model and the general recommended best practices, allowing opportunities for improving the model. Conducting a similar study with service providers may provide insight on the ability to defend PSAPs against TDoS/DDoS attacks specifically.

6.3 Conclusion

PSAPs are an integral component of the emergency services system in the United States. Disruption of the PSAPs can result in loss of lives. Cyberattacks of PSAPs is a real problem today, as over 40% of PSAPs and critical infrastructure sectors have already experienced them (Seals, 2017). Recognizing the near-term risk, the Department of Homeland Security, Science, and Technology Directorate (S&T) funded two research programs to harden defenses against DDoS and TDoS attacks. Whether PSAPs upgrade their current systems to include the NG911 concept or a major infrastructure upgrade is made, security measures are needed across all PSAPs. Implementation of this model is applicable regardless of the status of the PSAP infrastructure. Our security model captures the recommended cybersecurity best practices for protecting the PSAPs in the near term.

References

- 911.gov. (2016). *Next Generation 911*. U.S. National Highway Traffic and Safety Administration. Retrieved on November 28, 2017 from https://www.911.gov/issue_nextgeneration911.html
- APCO International. (2017). *Broadband implications for the PSAP: Analysing the future of emergency communications*. APCO International. <https://www.apcointl.org/download/apco-p43-report?wpdmdl=5984&ind=0>
- CISA. (2021). *Next Generation 911 Publications*. Retrieved on November 27, 2021 from <https://www.cisa.gov/publication/next-generation-911>

- CISA-Transition. (2021). Transition to Next Generation 911. Retrieved on November 26, 2021 from <https://www.cisa.gov/safecom/next-generation-911>
- FCC. (2017). *911 and E911 services*. Federal Communications Commission <https://www.fcc.gov/general/9-1-1-and-e9-1-1-services>
- Felicity Hansson, Sinead Keeney, and Hugh McKenna (2000). Research Guidelines for Delphi Survey Technique. *Journal of Advanced Nursing*, 32(4), 1008-1015.
- Guri, M., Mirsky, Y., & Elovici, Y. (2016). *911 DDoS: Threat, analysis and mitigation*. Ben-Gurion University of the Negev Cyber-Security Research Center. <https://arxiv.org/ftp/arxiv/papers/1609/1609.02353.pdf>
- Iqbal, S., & Papon-Young, L. (2009). *The Delphi method*. Retrieved July 28, 2018 from <https://thepsychologist.bps.org.uk/volume-22/edition-7/delphi-method>
- Jackson, D. (2017). APCO identifies potential NJ911 impacts, challenges in Project 43 report. *Urgent Communications*. Littleton.
- Kang, M., Gilgor, V., & Sekar, V. (2017). Defending against evolving DDoS attacks: A case study using link flooding incidents. *Springer International Publishing*, (24), 47-57.
- Kong, M., Kim, N., & Lee, G. (2017). AHP and Delphi method-based attack-capability evaluation methodology. *International Information Institute (Tokyo). Information*, 20(9A), 6497-6509.
- Motorola. (2019). Secure Next Generation Systems and Secure the PSAP. Retrieved on November 26, 2021 from <https://www.vestapublicsafety.com/pdf/Secure-the-PSAP-Whitepaper.pdf>
- NENA. (2017). *NENA master glossary*. Retrieved March 7, 2018 from https://cdn.ymaws.com/www.nena.org/resource/resmgr/standards/NENA-ADM-000.22-2018_FINAL_2.pdf
- NENA. (2017). *911 statistics*. Retrieved November 26, 2021 from <https://www.nena.org/?page=911Statistics>
- NENA. (2017). *NENA master glossary of 911 terminology; NENA-ADM-000.21-2017*. Retrieved November 28, 2017 from https://c.ymcdn.com/sites/www.nena.org/resource/resmgr/standards/NENA-ADM-000.21-2017_FINAL_2.pdf?hhSearchTerms=%22definition%22
- NVivo. (n.d.). *NVivo*. Retrieved January 5, 2019 from <https://www.qsrinternational.com/NVivo/home>
- SAFECOM. (2019). Cyber Risks to Next Generation 911. Retrieved on November 27, 2021 from <https://www.cisa.gov/sites/default/files/publications/NG911%20Cybersecurity%20Primer.pdf>
- Seals, T. (2017, March 29). *40% of ICS, Critical infrastructure targeted by cyberattacks*. Retrieved October 19, 2017 from <https://www.infosecurity-magazine.com/news/40-of-ics-critical-infrastructure/>
- Seppänen, H., Luukkala, P., Zhang, Z., Torkki, P., & Virrantaus, K. (2018). Critical infrastructure vulnerability – A method for identifying the infrastructure service failure interdependencies. *International Journal of Critical Infrastructure Protection*, 22, 25-38. doi:10.1016/j.ijcip.2018.05.002.
- Skulmoski, J., Hartman, G. T., & Krahn, J. (2007). The Delphi method for graduate research, *Journal of Information Technology Education: Research*, 6, 1-21. doi:10.28945/199.
- Turoff, M., Bañuls, V. A., Plotnick, L., Hilitz, S. R., & Ramírez de la Hueraga, Miguel. (2016). A collaborative dynamic scenario model for the interaction of critical infrastructures. *Futures*, 84, 23-42. doi:10.106/j.futures.2016.09.003.
- USA Patriot Act. (2001). Pub. Law 107-56, 107th Congress.
- White House. (2013, February 12). *Presidential Policy Directive – Critical infrastructure security and resilience*. Retrieved January 24, 2018 from <https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil>

Increasing Industry Profitability and Cyber Hygiene Utilizing Awareness Progression Methods

John Theborge, Mark Reith and Wayne Henry

Air Force Institute of Technology, Wright-Patterson AFB, USA

John.theborge@afit.edu

Mark.Reith@afit.edu

Wayne.Henry@afit.edu

Author Note

The views expressed are those of the authors and do not reflect the official policy or position of the US Air Force, Department of Defense, or the US Government.

Abstract: Securing critical networks and systems through proper cyber hygiene is a constant battle. Businesses spend a significant amount of time and money implementing cybersecurity mechanisms. However, businesses do not always see the cost-benefit from paying for proper cyber hygiene mechanisms, given the inevitability and persistence of cyber threats. This research explores potential financial incentives for businesses to improve their cyber hygiene awareness. Past anti-smoking and climate change awareness campaigns are compared to support a new cyber hygiene awareness campaign. By investigating the effectiveness of the incentive methods used by these awareness campaigns, this work proposes adopting similar incentive methods to improve cyber hygiene awareness.

Keywords: Cyber Hygiene, Profitability, Awareness Campaigns, Climate Change, Anti-Smoking

1. Introduction

As the Internet and the interconnection of systems grows, the likeliness for cyber incidents to occur grows with it. Cyber incidents frequently exploit the poor cyber hygiene of the Internet's users. Cyber hygiene refers to best practices of users to ensure security of computers, cyber-physical systems, and data (Brook, 2020). The processes and mechanisms required to obtain proper cyber hygiene have always been expensive. The extent of these costs is shown in the global spending on cybersecurity, which was \$123 billion USD in 2020 and is projected to be \$133 billion in 2022 (Shackleton, 2021). The remediation costs associated with cyberattacks can also be significant. In 2020, global cyber-attacks cost government and industry \$1 trillion, averaging \$3.9 million per incident. Deloitte and the Financial Services Information Sharing and Analysis Center (FSISAC) performed a study that surveyed 97 major institutions that had annual revenues between \$500 million and \$2 billion. This report discovered that the companies had increased their cybersecurity budget from 0.34% of their total revenue in 2019 to 0.48% in 2020 (Bernard, 2020). This shift in funding is a reactionary effort from industry to combat the growing number of total breaches between 2018 and 2019. In 2018, approximately 1.257 billion data breaches occurred with 472 million individual records exposed, compared to 2019 with 1.473 billion data breaches and 164 million individual records. (Johnson, 2021). In 2020, the total data breaches dropped to 1 billion, with nearly 156 million individual records exposed. Despite the reduced number of breaches in 2020, a reactionary force is still a losing force in the cyber domain. Allowing companies to remain in a reactionary position in the cyber domain allows a higher level of risk acceptance.

This is not the first-time companies were required to adapt to other major hygiene awareness campaigns. When smoking became a global health concern, companies had to incentivize their employees not to smoke, as it was costing them more due to their employees' decreasing health (Baker et al, 2017; Sammer, 2019). A similar outcome occurred with climate change when companies had to reduce their total emissions. Companies were able to reduce their emissions by utilizing incentives that helped offset the cost of changing to reduced emission processes (EDF, 2018).

This paper investigates the similarities that exist between the anti-smoking, climate change, and cyber hygiene awareness campaigns. We then discuss the financial incentives used to progress the anti-smoking and climate change awareness campaigns. This paper concludes by proposing the adoption of similar financial incentive methods to progress cyber hygiene awareness and reduce the cost of safeguarding assets from persistent cyber threats.

2. Comparing Anti-Smoking and the Cyber Hygiene Awareness Campaigns

Smoking kills more than 480,000 people per year in the United States alone, which includes more than 41,000 deaths from second-hand smoking (CDC, 2021). The general population was unaware of the negative health effects cigarettes had when they gained in popularity in the 20th century (Lorgat, 2020). The trend of smoking grew rapidly. It was not unusual to see characters smoking in children's books or famous actors smoking on television. In the 1960s, the Surgeon General publicly recognized the overwhelming evidence that displayed the negative impacts of smoking, such as lung cancer and bronchitis. This recognition came after 45% of Americans were already smoking. The percentage of Americans that smoked has gradually dropped since then. In 2015, 15.2% of Americans smoked (Blakemore, 2015). Despite the 30% decrease of total American smokers since the 1960s, the damage of cigarettes continues to cost over \$255 billion in direct medical care for adults and \$156 billion in lost productivity each year (CDC, 2021). These costs are associated with smokers and their employers having to pay higher health insurance rates due to their higher risk of illness (Baker et al, 2017; Sammer, 2019). The popularity of cigarettes demonstrates the tendency for society to adopt the latest trends based on its general acceptance without concern for the negative long-term impacts it could have on their individual health and the health of those around them.

The Internet became popular in a similar fashion as smoking did. Unfortunately, the popularity of the Internet has led to poor cyber hygiene of its users. In 1967, the concept of ARPANET (The Advanced Research Projects Agency Network) was created by the Department of Defense (DoD) as the first packet-switched network that allowed for the rapid dissemination of information from one network node to another (Leiner et al, 1997). In 1972, ARPANET was successfully demonstrated, sparking the rapid adoption and creation of communication applications around the world. In 1983, this rapid growth of networks and the seemingly endless capabilities they would provide soon evolved into the Internet. Today, the Internet now has 4.88 billion active users (Johnson, 2021; Leiner et al, 1997).

Very few cared to recognize the threats that would develop with the mass adoption of the Internet. Due to this lack of recognition, malicious actors were able to make these threats a reality early on. In 1986, a German hacker by the name of Marcus Hess used an Internet gateway in Berkeley, CA, to gain access to the ARPANET and MILNET (Chadd, 2020). Hess was able to hack into 400 military computers. Hess had intentions to sell the information to the KGB, who were the main security agency for the Soviet Union. In the 1990s, polymorphic viruses made their first appearance. Polymorphic viruses avoided detection by mutating as it traversed systems. These types of attacks avoided commonly used signature-based antivirus software, leaving no proper methods of detecting and preventing these types of attacks. In 1999, the Internet saw the aggressive propagation of the Melissa virus. The virus propagated via email utilizing the users' email address book of the infected machine.

As the Internet developed, computers became more interconnected. It was near impossible to stop or even deter malicious actors as the threat space was growing exponentially worldwide. As a result, there was a rapid decline of cyber hygiene across companies and those who had adopted the Internet. The capabilities the Internet provided led to its increase in popularity. The Internet's popularity pushed the concern for its lack of security into the background (Timberg, 2015). Today, we continue to see the lack of emphasis on cyber hygiene, considering the number of security breaches that are still occurring. In 2005, there were 157 million data breaches in the United States and 66.9 million individual records were exposed (Johnson, 2021). In 2020, these incidents increased by over 537%, with 1 billion data breaches in the United States, exposing approximately 156 million individual records.

Both the Internet and smoking rapidly grew through popularity without concern for the long-term negative effects they would have. Poor cyber hygiene causes hardship to companies and individuals through expensive remediation costs for cyber incidents and stolen individual records. Stolen identities through the exposure of these records cause direct hardship to individuals and their families. This hardship is caused through impacted credit scores, credit card fraud, and other financial burdens. Similarly, smoking causes hardship to the smoker's employer through increased health coverage costs and reduced productivity. Smokers also see financial burdens and health issues for themselves and their families. The increased likeliness of illness leads to higher health premiums and an increase in medical costs. Second-hand smoke can also impact families and their health. Given the similarities between smoking and poor cyber hygiene, we will examine the incentives used in anti-smoking awareness and how they could be utilized to progress cyber hygiene.

3. Comparing Climate Change and the Cyber Hygiene Awareness Campaigns

During the 18th and 19th centuries, the world saw the Industrial Revolution. The Industrial Revolution led to the rise of global production and new manufacturing processes around the world (CPW, 2021). This rise in global production and manufacturing led to the development of emissions. The pursuit of manufacturing led to the industrialization of once-rural areas into factory-filled cities. With industrialization came the dependency on fossil fuels such as coal, natural gas, and oil. The use of these fossil fuels resulted in about 65% of the world's greenhouse emissions through the powering of transportation, heating, and electricity generation. The movement for climate change came after recognizing the negative impacts that the emission processes were having on the planet. The climate change awareness campaign has made several major leaps forward in governance and policy. In 1997, the Kyoto Protocol was introduced as one of the first international policies for reducing emissions around the world. In 2008, it was accepted by 181 countries (Zhou et al, 2018). The Kyoto Protocol outlined the obligations developed countries needed to follow to reduce greenhouse gas emissions and outlined the means for the countries to accomplish this. This included emissions trading and carbon credits. The Kyoto protocol also introduced "Clean Development Mechanisms," which provided carbon or tax credits for investing in technology and infrastructure that reduced emissions. The Paris Agreement was another major international agreement created to ensure the world is progressing the climate change campaign. Activated in 2016, it introduced the idea of international carbon trading markets through its article 6 (EDF, 2021). Additionally, it outlined the objective of reducing the rise in mean global temperature to below 2-degrees Celsius (UNFCCC, 2021).

When comparing climate change and cyber hygiene, we recognize that cyber hygiene was not at the forefront through the rapid adoption and popularity of the Internet. Similarly, the negative impact of emissions on our climate was also not at the forefront through the adoption and popularity of industrialization. In the climate change awareness campaign, we saw the implementation of international policies due to the high level of emissions worldwide. In the same vein, cyber hygiene is also an international concern. We saw the acknowledgment of cyber being an international concern in the United States International Strategy for Cyberspace (USISC) under the Obama administration (Congress, 2011). This cyber strategy addressed the need for a stable cyberspace and the critical role it plays in the success of the global economy. The strategy also outlined the need to develop international norms of behavior in cyberspace, and international cybersecurity capacity building. The policy also mentioned the necessary frameworks to stabilize cyberspace. These frameworks emphasized the protection of critical infrastructure. Like the efforts of the Kyoto Protocol and the Paris Agreement combatting carbon emissions worldwide, the USISC stresses the need for a global effort on cyber hygiene by recognizing worldwide cyber threats and risks. Like climate change, cyber hygiene is a global issue. Both efforts require key world players to work together to combat the diminishing health of our planet and the cyber hygiene of Internet users.

4. Incentives in Climate Change and Anti-Smoking Awareness Campaigns

Some of the incentive methods currently being used to progress the climate change and anti-smoking awareness campaigns are as follows:

4.1 Climate Change Emissions Trading

Emissions trading, commonly referred to as "cap-and-trade" programs, provides economic incentives to companies to reduce their greenhouse gas emissions. One of the major US-based programs is the Regional Greenhouse Gas Initiative (RGGI). This program was the first market-based initiative that involved companies in Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New York, Rhode Island, Vermont, and Virginia (Ceres, 2020; RGGI, 2021). The program operates by having an authority provide a maximum cap of greenhouse gas pollution per region. The cap is then split into individual limits and issued to companies. The limit indicates the total greenhouse gas pollution the company can produce in a three-year timeframe. If a company is below their maximum emission limit, they can sell their unused portion to a buyer company, turning their reduced emission expenses into profit. Alternatively, if a company knows they will exceed their limit, they can negotiate to buy another company's unused portion. This prevents the company from paying steep fines for going over their allotted limit. If companies refuse to implement the necessary processes and procedures, they will be forced to continue buying other companies' unused limit portions or go over their allowed limit. This will ultimately cost them more in the long term, given the cap for the region is lowered over time.

The cap-and-trade program forces companies to reduce their overall emissions and has yielded positive results. The decrease of emissions due to the lowering of the authorized cap over the years within the RGGI program can be seen in Figure 1. This graph shows the gradual decrease in emissions (per million tons) as the cap is lowered annually (Climatekenex, 2019; RGGI, 2021). This program has not only led to the reduction of greenhouse gas emissions but provided over \$416 million in revenue in 2020 alone. Since 2005, RGGI has collected \$3.8 billion in total revenue (ICAP, 2021). This revenue is returned to the states involved in the RGGI program, where it has been primarily invested in consumer benefit programs.

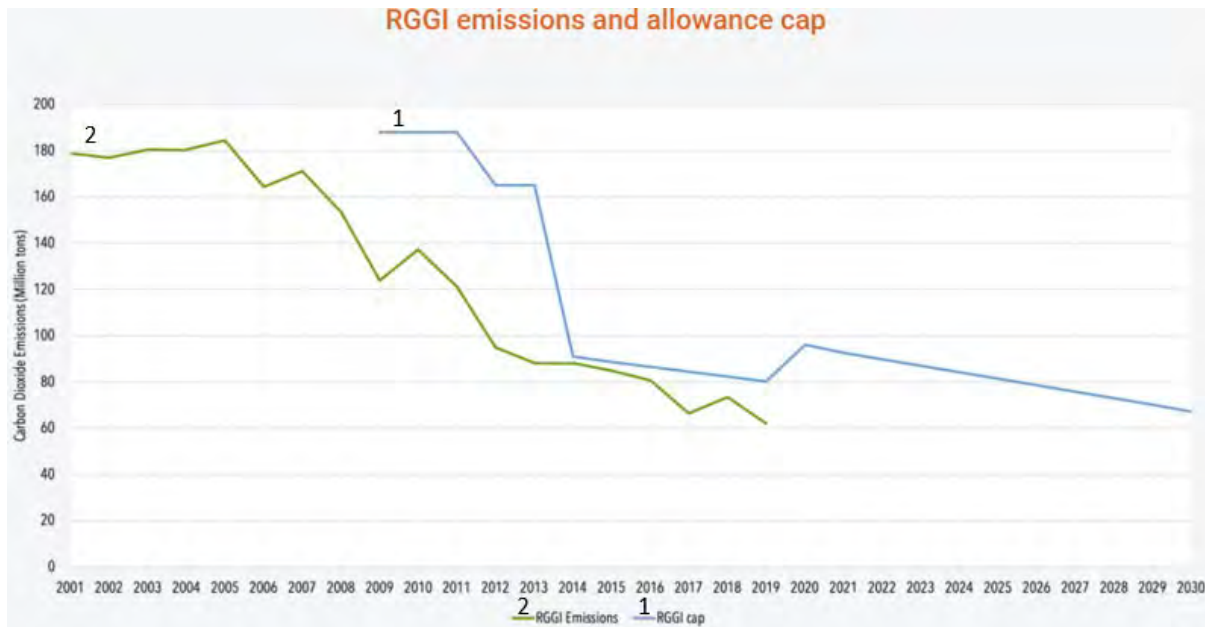


Figure 1: Total emissions against the lowered cap in RGGI (C2ES, 2020).

In the United States, California has implemented a similar cap-and-trade program called the “AB 32 Scoping Plan.” The AB 32 Plan has seen a 10% decrease in total emissions between 2013 and 2018 and aims at achieving 40% below 1990 emission levels by 2030 (EDF, 2016; 2018). Emissions trading has also proven to be successful outside of the United States. Europe has seen success with the European Union’s Emissions Trading System (EUETS), as well as China. China produces the largest emissions of greenhouse gases and implements cap-and-trade elements into their new emissions trading system. In 2018, they saw a 29% decrease in stationary structure emissions when compared to 2005 emission levels (EDF, 2018). Emissions trading has not proven its success internationally between countries. The Paris Agreement Article 6 introduced the concept of international emissions trading, but it becomes a more difficult incentive to implement across different countries given the international rules that must be in place and agreed upon by all parties. Recently approved in mid-November 2021, its international effectiveness has not been determined (Hedley, 2021).

4.2 Climate Change and Carbon Capture Credits

Carbon capture and sequestration is the process of injecting carbon oxides underground after they have been captured from the original emission source, such as a powerplant (Hasan et al, 2015). The power of carbon capture is immense. In an optimal nationwide design of carbon capture, utilization, and sequestration (CCUS), it is possible to capture over 1.5 Gt (gigatons) per year of carbon emissions. This requires the capture of 50% of stationary structure emissions. Figure 2 presents the total carbon emissions captured from stationary structures. As the reduction level increases from 60% to 80%, it yields a total of over 1.8 Gt/year and over 2.4 Gt/year, respectively.

Nationwide CCUS costs.

	CO ₂ reduction level		
	50%	60%	80%
Total stationary CO ₂ emission (Gt/year)	3.068	3.068	3.068
Total CO ₂ captured (Gt/year)	1.534	1.841	2.454
Total flue gas dehydration cost (billion dollars/year)	15.677	18.815	25.080
Total capture & compression cost (billion dollars/year)	38.238	46.823	70.906
Total transportation cost (billion dollars/year)	2.551	3.870	7.636
Total injection cost (billion dollars/year)	1.627	2.001	2.994
Total CCUS cost (billion dollars/year)	58.093	71.509	106.616
Total CCUS revenue (billion dollars/year)	3.43	3.52	3.60
Overall CCUS cost (billion dollars/year)	54.663	67.989	103.016
Overall CCUS cost (dollars/ton CO ₂ captured)	35.634	36.930	43.445

Figure 2: Total costs for an optimal CCSU supply chain network in the United States (Hasan et al, 2015).

The total carbon captured demonstrates the strength of this technology, while the cost shows the extreme prices associated with using it. To successfully capture 50% of stationary structure emissions, it would cost over \$54 billion per year (Hasan et al, 2015). This is where carbon or tax credits are used. Companies and individuals who invest in carbon capture projects or technology can receive tax or carbon credits per ton of carbon successfully captured and sequestered. These tax credits can range from \$35 to \$50 per ton of carbon. To receive this benefit, the total carbon captured must range between 25,000 and 500,000 metric tons. This could yield anywhere between \$875 thousand to \$25 million worth of credits. The range specified is dependent on how much the carbon emission facility produces annually (CRS, 2021). A limiting factor of the effectiveness of carbon-capturing is the immense cost that is associated with it. The technology is expensive to develop, and the cost is preventing an ideal supply chain network in the United States from coming to fruition (Baylin-Stern, 2021).

4.3 Anti-Smoking Incentives

As previously highlighted, the damage of cigarettes continues to cost \$255 billion in direct medical care for adults and \$156 billion in lost productivity (CDC, 2021). The loss of productivity is an indirect cost to a company because it is caused by the increased risk of illness from smoking, leading to absenteeism. In 2017, Pfizer surveyed 75,000 individuals across the United States, EU, and China. They found that active smokers had 28% more absenteeism over a seven-day span than those who had never smoked or had quit (Baker et al, 2017). Companies recognized smokers were costing them more annually to provide adequate healthcare coverage than a non-smoker. To encourage smokers to quit, surcharges and incentives were introduced. In 2004, a study involving General Electric found that offering a cash incentive of \$750 yielded almost three-times as many people quitting as those without the incentive. This increased the quit rate from 5.0% to 14.7%. This ultimately led General Electric to provide an incentive program for all United States employees. This study was later published by the New England Journal of Medicine and became a widely adopted approach by governments, employers, and insurers to progress the anti-smoking campaign further (Halpern et al, 2015).

A similar study with CVS employees and their families tested the effectiveness of different variations of incentive programs. This study which had 2,538 participants focused on two main approaches. The first approach focused on an individual-oriented incentive program that offered either a reward for the individual or a reward with an initial deposit by the participant. The second approach was identical to the first, except it was group-oriented with 6 participants per group (Halpern et al, 2015). The reward offered was \$800 for smoking cessation past six months. The deposit study entailed a refundable deposit of \$150 and \$650 in a reward payment. The study results showed that 90% of participants accepted the reward-based program. Only 13.7% of participants accepted the deposit-based program. Overall, the reward-based programs had a higher cessation rate of 15.7% compared to 10.2% from the deposit-based program. The difference in success rates between individual-oriented and group-oriented incentive programs was considered negligible, with success rates of 12.1% and 13.7%, respectively. Additionally, the individual reward-based program saw a 9.7% higher smoking cessation rate compared to those under usual care at 6%. Figure 3 represents the outcomes of the CVS study. The difference in the success of both individual and group-based rewards can be seen compared to the deposit-based incentive. One of the major highlights from Figure 3 is the cessation rate through 12 months, where both reward programs maintained the highest percentage of smoking cessation.

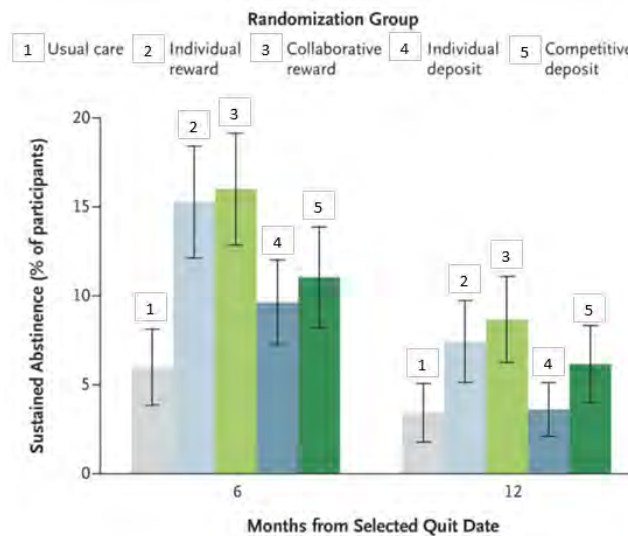


Figure 3: Smoking cessation success rates for reward-based and deposit-based incentive programs (Halpern et al, 2015).

These incentive-based programs may cost companies more in the short term due to the upfront costs of the rewards. Although, it is estimated that having these types of programs will save an employer anywhere between \$150 and \$540 per additional non-smoking employee (Ekpu & Brown, 2015). This has been a widely adopted incentive program because of its success and has ultimately turned a profit for companies over a multi-year timeframe. It is important to note that the cessation rate into 12 months dropped by about 8% compared to the six-month interval at 15.7%. Although the six-month timeframe is relatively impressive, the effectiveness of the program does not seem as impressive over the longer 12-month period (Halpern et al, 2015).

5. Proposal

The anti-smoking and climate change awareness campaigns used multiple methods to incentivize companies to progress their overall awareness. Again, these incentive methods included cash rewards, tax and carbon credits, and trading. Utilizing the incentive methods introduced in climate change, we have a foundation of a cyber risk trading concept. Instead of an emissions limit for climate change over three years, we have an allowable cyber risk limit for a company over a similar timeframe. Cyber risk in a company is measured through identifying critical infrastructure and assets, the impact of losing those assets, and identifying cyber threats and their applicability to the respective company (Dumont, 2020). This allows cyber risk to be a quantifiable metric. This measurement of risk can be used by a government-based authority to determine the maximum limit of cyber risk per a given region. Regions would consist of different companies with different lines of business. For example, the cyber risk of health institutions will be different from the risks pertaining to financial institutions; therefore, it would be unfair to categorize these types of companies under the same cyber risk allowance. Additionally, the authority will lower the risk cap annually, forcing companies to reduce their cyber risk, ultimately improving their cyber hygiene over time. Companies can trade their unused risk limits to other companies who need them. This would assist in offsetting the cost of implementing adequate cyber hygiene processes and mechanisms. This concept also introduces the potential of utilizing an existing framework, the Cybersecurity Maturity Model Certification (CMMC). The CMMC is a framework that assesses an organization's implementation of cybersecurity practices (Dumont, 2020). These assessments are done by a certified 3rd party organization. The CMMC framework is expected to be implemented across the DoD over the next five years, and its use of certified 3rd party assessment organizations could be leveraged to assess the cyber risk of the companies in the risk trading concept (Dumont, 2020). Forcing companies to reduce their risk over time while financially incentivizing them through the trading concept would progress the cyber hygiene of the companies involved.

Continuing with climate change incentives, we have carbon capture with investment benefits. The extreme costs associated with carbon-capturing technology are akin to the extreme costs in implementing proper cyber hygiene. The use of outside investors could prove beneficial to the progression of cyber hygiene, just like it has with carbon capturing. Instead of carbon capture, we have a concept of cyber risk capturing. The concept of cyber risk capturing is to provide an investor company with risk credits or tax credits that choose to invest in the progression of cyber hygiene in other companies. The investing companies can use the risk credits they receive

to offset their own cyber risk limit. Risk capturing and its usability can be seen with an example of an increase in ransomware attacks towards health institutions. The authority overseeing the health institutions would lower their risk limits to combat the increase in ransomware. In this scenario, an outside company not directly targeted by the increase in ransomware attacks can choose to invest in the progression of cyber hygiene in the health institution. Not only will the outside investment aid in increasing the cyber hygiene of the health institution and lower their risk limit, but the investing company can use the credits they would receive to offset their own cyber risk limit as well.

Moving to anti-smoking awareness incentives, we have a concept for positive cyber hygiene incentives. Companies will provide incentive funds to their employees who collectively increase their cyber hygiene on an annual basis. Companies would gauge an increase in cyber hygiene across their employees through the application of cyber hygiene best practices. The use of these best practices would reduce the number of incidents that occur through the end-user. The funds will be restricted to being spent solely on employee improvement programs. The goal is to incentivize the employees to apply what they learn in their cyber hygiene training. Doing so will directly increase the company's cyber hygiene and ultimately save them money from the reduced number of cyber incidents.

6. Conclusion

This paper recognizes similarities between the anti-smoking, climate change, and cyber hygiene awareness efforts. Through these similarities, we identified that cyber hygiene, personal hygiene, and the health of our planet rapidly declined through the popularity of the Internet, cigarettes, and industrialization, respectively. Society has shown a tendency to adopt the latest trends based on its public acceptance and not think twice about the negative long-term impacts these trends could have. It is important to recognize that the previous awareness campaigns are just as much a social issue as they are a technical issue. This enables the possibility of adopting progression methods other awareness efforts have used to progress other seemingly unrelated awareness campaigns.

The smoking and climate change awareness campaigns recognized that incentives were needed to get companies to progress their awareness, which ultimately yielded positive results in their respective efforts. Cyber hygiene is in a similar position. Companies need to be incentivized to properly safeguard their assets and personnel due to the increasing costs of proper cyber hygiene mechanisms and the persistence of cyber threats. Adopting a cap-and-trade program for cyber risk would reduce the cyber risk companies are accepting. This would increase their overall cyber hygiene and would prove to be financially beneficial. Similarly, the adoption of carbon or tax credits for cyber risk could reduce a company's overall risk, which would increase their cyber hygiene as well.

References

- Baker, C. L. et al., 2017. Benefits of quitting smoking on work productivity and activity impairment in the United States, the European Union and China. *International Journal of Clinical Practice*, 1.71(1).
- Baylin-Stern, A., 2021. Is carbon capture too expensive?. <https://www.iea.org/commentaries/is-carbon-capture-too-expensive>.
- Bernard, J., 2020. *Reshaping the cybersecurity landscape*, s.l.: s.n.
- Blakemore, E., 2015. In 1965, 45 percent of Americans smoked, today it's only 15 percent.. <https://www.smithsonianmag.com/smart-news/only-15-percent-americans-smoke-180956504>.
- Brook, C., 2020. What is Cyber Hygiene? A definition of cyber hygiene, benefits, best practices, and more. <https://digitalguardian.com/blog/what-cyber-hygiene-definition-cyber-hygiene-benefits-best-practices-and-more>.
- C2ES, 2020. *Regional Greenhouse Gas Initiative (RGGI)*. [Online]
Available at: <https://www.c2es.org/content/regional-greenhouse-gas-initiative-rggi/>
- CDC, 2021. Tobacco Fast facts. Centers for Disease Control and Prevention.
https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm.
- Ceres, 2020. Ceres Report.
<https://www.ceres.org/sites/default/files/Fact%20Sheets%20or%20misc%20files/March%2015%20S177%20Letter%201.pdf>.
- Chadd, K., 2020. The history of Cybersecurity. <https://blog.avast.com/history-of-cybersecurity-avast..>
- Climatenexus, 2019. Emissions trading. [https://climatenexus.org/climate-change-us/politics-and-policy/emissions-trading/..](https://climatenexus.org/climate-change-us/politics-and-policy/emissions-trading/)
- Congress, U. S., 2011. International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World..
- CPW, 2021. The warming effects of the Industrial Revolution - global temperatures. <https://www.climate-policy-watcher.org/global-temperatures/the-warming-effects-of-the-industrial-revolution.html..>

- CRS, 2021. The tax credit for carbon sequestration (section 45Q).. <https://sgp.fas.org/crs/misc/IF11455.pdf>.
- Dumont, C., 2020. CMMC - risk assessment. <https://www.tenable.com/sc-dashboards/cmmc-risk-assessment>.
- EDF, 2016. California leads fight to curb climate change.. <https://www.edf.org/climate/california-leads-fight-curb-climate-change>.
- EDF, 2018. How cap and trade works.. <https://www.edf.org/climate/how-cap-and-trade-works>.
- EDF, 2021. Cop 26: Implementing Article 6 of the Paris Agreement.. <https://www.edf.org/climate/implementing-paris-climate-agreement>.
- Ekpu, V. U. & Brown, A. K., 2015. The Economic Impact of Smoking and of Reducing Smoking Prevalence: Review of Evidence. *Tobacco Use Insights*, 1.Volume 8.
- Greig, J., 2020. 66% of companies say it would take 5 or more days to fully recover from a ransomware attack ransom not paid.. from <https://www.techrepublic.com/article/66-of-companies-say-it-would-take-5-or-more-days-to-fully-recover-from-a-ransomware-attack-ransom-not-paid>.
- Halpern, S. D. et al., 2015. Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation. *New England Journal of Medicine*, 5.372(22).
- Hasan, M. F., First, E. L., Boukouvala, F. & Floudas, C. A., 2015. A multi-scale framework for CO2 capture, utilization, and sequestration: CCUS and CCU. *Computers & Chemical Engineering*, 10.Volume 81.
- Hedley, A., 2021. Paris agreement article 6 success - A long time coming. <https://www.lexology.com/library/detail.aspx?g=75d8c877-c351-494d-a14b-44ebe0f899e4>.
- ICAP, 2021. USA - Regional Greenhouse Gas Initiative (RGGI). https://icapcarbonaction.com/en/?option=com_etsmap&task=export&format=pdf&layout=list&systems%5B%5D=50.
- Johnson, J., 2021. Internet users in the world 2021. from <https://www.statista.com/statistics/617136/digital-population-worldwide>.
- Johnson, J., 2021. U.S. data breaches and exposed records 2020.. <https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/>.
- Leiner, B. M. et al., 1997. The past and future history of the Internet. *Communications of the ACM*, 2.40(2).
- Lorgat, I., 2020. Smoking: A 100-year story that doesn't end here. Smoking: a 100-Year Story That Doesn't End Here.. <https://www.rgare.com/knowledge-center/media/articles/smoking-a-100-year-story-that-doesn-t-end-here>.
- Murphy, P., 2020. To defend forward, US cyber strategy demands a cohesive vision for information operations.. <https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/2420176/to-defend-forward-us-cyber-strategy-demands-a-cohesive-vision-for-information-o/>.
- Pearlson, K., 2021. Cyberattacks are inevitable. is your company prepared?. <https://hbr.org/2021/03/cyberattacks-are-inevitable-is-your-company-prepared>.
- RGGI, 2021. About the Regional Greenhouse Gas Initiative. https://www.rggi.org/sites/default/files/Uploads/Fact%20Sheets/RGGI_101_Factsheet.pdf.
- Sammer, J., 2019. Employer incentives encourage employees to quit smoking. <https://www.shrm.org/hr-today/news/hr-magazine/1118/pages/employer-incentives-encourage-employees-to-quit-smoking.aspx>.
- Shackleton, T., 2021. A cost-benefit analysis approach to cyber security. <https://www.6dg.co.uk/blog/cost-benefit-approach-to-cyber-security/>.
- Timberg, C., 2015. The real story of how the internet became so vulnerable. <https://www.washingtonpost.com/sf/business/2015/05/30/net-of-insecurity-part-1/>.
- UNFCCC, 2021. Key aspects of the Paris Agreement. <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement/key-aspects-of-the-paris-agreement>.
- Zhou, Y., Hu, F. & Zhou, Z., 2018. Pricing decisions and social welfare in a supply chain with multiple competing retailers and carbon tax policy. *Journal of Cleaner Production*, 7.Volume 190.

Identifying Adversaries' Signatures Using Knowledge Representations of Cyberattack Techniques on Cloud Infrastructure

Gilliam van der Merwe, Christian Muller, Wilhelm van der Merwe and Dewald Blaauw
Department of Information Sciences, Stellenbosch University, Cape Town, South Africa

gvdwm.work@gmail.com

chrismullerpsa@gmail.com

dwilhelmvdm@gmail.com

dnblaauw@sun.ac.za

Abstract: Advanced Persistent Threats (APTs) have increased in parallel to growing cloud infrastructure and cloud Software-as-a-Service (SaaS) needs, exposing new vulnerabilities within the cloud environment. Moreover, APT groups are becoming more sophisticated and organised which needs to be addressed by the research community to enable faster response and more importantly, prevent threats within the domain. The MITRE ATT&CK Cloud framework offers one of the leading structured inventories within this context. Our research is to expose patterns and signatures of a select group of APT's on the MITRE Cloud Framework by using Formal Concept Analysis (FCA) to construct a "lattice graph" and an ontology. The goal is to develop a better conceptualisation of the MITRE ATT&CK Cloud Matrix framework for cyber security experts to be able to proactively act upon adversary techniques. The MITRE ATT&CK framework was retrieved, cleaned, and pre-processed to construct the lattice and ontology using data cleaning methods, FCA tools such as Concept Explorer, and the Web Ontology Language (OWL), with additional symbolic reasoning and inference generation. This resulted in knowledge representations/graphs, which are highly efficient representations of this knowledge field. The underlying linkages between techniques and targets specific to those APTs are further exposed and enriched and presented visually and integrated into the ontology. The ontology gives formalisation to associations and implications between techniques, tactics, and APTs – enabling cyber security practitioners to forecast potential targets and techniques based on their scenario, but also to attribute certain technique patterns and signatures to individual APTs. Cyber security practitioners can query from this knowledge graph and formulate strategic proactive measures. From these findings, the applications and constraints of the APTs' cyber-attack techniques and their associated patterns were determined. The findings provide a guideline for future additional research in the field of AI knowledge representation in cybersecurity, as well as highlighting certain limitations in this field of research.

Keywords: Cloud Infrastructure, Cybersecurity, Knowledge Representation, Advanced Persistent Threats

1. Introduction

Identifying APTs through signatures derived from utilizing a knowledge representation of cyberattack techniques on cloud infrastructure uncovers previously unknown information which could in turn prevent losses and transcend the knowledge of the cyber security community. The extended use of existing information given on the MITRE ATT&CK cloud matrix framework aims to expose underlying patterns and signatures of techniques that are of a similar fashion and nature. The MITRE ATT&CK Cloud matrix framework, an underlying branch of the MITRE knowledge base, lays out several adversary techniques formulated from actual world-wide attacks on Cloud infrastructure. An APT is a clandestine cyber-attack found on a network with the goal to exfiltrate and/or damage information. These attacks can be imposed by small to large groups, in some cases covert government -backed entities. The research team uses a Knowledge Representation in AI to highlight these underlying patterns which become apparent through this form of representation. The representation will pay particular attention to include correlations within the techniques used by six of these malicious groups exclusively, referred to as Advanced Persistent Threats (APT's), as to display the extensive potential value this research holds but without the burden of substantial complexity. Although current methods exist to detect APTs through their signatures, most of these methods are black box methods. This research uses KR to gain a deeper understanding of the relationships between attack techniques as well as threat actors (APTs) that make use of a certain set of techniques. By mapping the APTs in a KR the study hopes to predict what the next stage within a sequence of attacks will be. More specifically, the FCA implication and association sets from the KR will be utilised to generate a glass box model of these relationships in order to formulate a visual interpretation in the form of an ontology which enforces advanced comprehension, and in turn could allow for a faster response or proactive defence against these attacks.

In comment to the remainder of this paper, which structure reflects as follows: a presentation of background information which provides context surrounding Cloud Infrastructure and Cloud SaaS, MITRE ATT&CK Cloud matrix framework, and the formulation of an ontology enabled through FCA. Followed by an in -depth gradual

discussion of the data exploration and analysis methodology used to extract the sought-after and meaningful insights described within the outset of this research paper. In closing, research results and a description of possible future research is presented.

2. Background and Related works

2.1 Cloud Infrastructure and cloud SaaS

Cloud Infrastructure is a term encompassing the different components that are required for cloud computing. To be able to host varying cloud applications and services, specific infrastructure is required. Cloud infrastructure works by way of virtualisation. This is the process by which the processing capabilities of hardware are utilised for cloud computing, by means of software that allocates resources to environments that users can access anytime and scale to what the users require. (Red Hat, 2021)

2.2 MITRE ATT&CK framework

The MITRE ATT&CK framework is a publicly accessible trusted and standardised framework that provides a knowledge base of information on cyberattack tactics and techniques. These tactics and techniques are based on actual cyberattack examples. This knowledge base also contains information on the Advanced Persistent Threats (APTs) that utilise these tactics and techniques. This knowledge base serves as a basis for the development of cyber threat models and methodologies in various sectors both public and private (Strom, et al., 2020).

The MITRE ATT&CK Cloud Matrix is a more specialised alternative to the MITRE ATT&CK Enterprise framework as it focuses on adversary techniques in a Cloud infrastructure context.

2.3 Formal Concept Analysis

FCA is, according to Priss (2006), a method for data analysis, knowledge representation, and information management. In short, it allows us to hierarchically group objects according to their common attributes by using the mathematical theory of concepts and conceptual hierarchies, which forms the basis for FCA (Ganter & Obiedkov, 2016). The mathematical theory for FCA is beyond the scope of this paper but can be viewed in (Ganter & Obiedkov, 2016). Furthermore, FCA is based on the theory that a concept consists of two parts namely, its extension, which includes all the objects belonging to the concept, and its intension which consists of all attributes shared by these objects (Obitko, Snasel, & Smid, 2004). Additionally, formal concept lattices enable the visualisation of relationships between concepts, which in turn allows for a more intuitive manner to interact with the data (Kuznetsov & Watson, 2017).

In terms of FCA tools, various implementations exist (Priss, 2006), however, ConExp and Lattice Miner was best suited in relation to the objectives of this paper. Lattice Miner is an open-source Java tool that allows representation and manipulation of input data, lattices, implication sets, and association rules (Bertet, Borchmann, Cellier, & Ferre, 2007). ConExp works in a similar fashion, however, the tool works better for attribute exploration (Priss, 2006).

Association rules allow relationships within the concept lattices to be categorised depending on their confidence and support values. Support signifies the frequency of a relationship occurring, whereas confidence represents the probability of relationships to co-occur (Pasquier, 2000).

2.4 Ontology Engineering

Semantic technologies allow computers to interpret data in order to derive meaning thereof, this is made possible using formal semantics by defining concepts and how they relate. To allow computers to interpret a knowledge base an ontology is utilised to represent the said knowledge base (Gronberg, 2019). More specifically, ontologies model and define a domains entities, classes, properties and relationships (Gruber, 2009).

By using an ontology, knowledge exchange is facilitated by implementing a shared vocabulary. Moreover, ontologies can be used in conjunction with inference engines to enable reasoning, it is also used for user interface, database, and application components (Gronberg, 2019). The Protégé ontology editor is an open-source tool that enables ontologies to be built with all the necessary implementations and it conforms to the OWL 2 Web Ontology Language and RDF standards from the World Wide Web Consortium (Mussen, 2015). Protégé also allows the use of inference engines to enable reasoning.

2.5 Related Work

Recent research on the application of Knowledge Representation, and more specifically Formal Concept Analysis (FCA), on the MITRE ATT&CK Framework showed promising results for the Cyber Security domain. Watson & Watson (2020) explored the use of FCA as a means for exploratory data science focusing on the relationship between adversary attack techniques and procedures of the mobile/IOS device subset in the ATT&CK framework. This approach enabled them to not only detect which components are shared between threats but also the possible identification of attribution signatures, prediction of new threats and enabling security practitioners to take precautionary measures.

Similarly, Singels, et al. (2020), applied FCA to the Industrial Control Systems (ICS) ATT&CK framework to enable 'earlier and strategic detection of cyberthreats'. They went further by encoding the FCA results to a formal ontology, in this case using the Web Ontology Language (OWL), allowing machine interpretation and semantic web reasoning. Through this, security practitioners are able to query the FCA results facilitating the discovery of knowledge.

3. Problem statement and research objectives

Although current methods exist to detect APTs through their signatures, most of these methods are black box methods. This research aims to use KR to gain a deeper understanding of the relationships between attack techniques as well as threat actors (APTs) that make use of a certain set of techniques. By mapping the APTs in a KR the study hopes to predict what the next stage within a sequence of attacks will be. More specifically, the FCA implication and association sets from the KR will be utilized to generate a glass box model of these relationships to see if a visual interpretation could aid in a better understanding, and in turn allow for a faster response or proactive defence against these attacks.

Questions that this research aims to address:

1. Are Knowledge Representations an accurate and efficient method of illustrating cyber-attack techniques?
2. What implications can be drawn from the knowledge representations of cyberattack techniques?
3. What insight can be gained from the researched cyberattack toolkits?
4. Is it possible to derive (*useful and identifying*) patterns and signatures of APTs from mapped data of cyber-attack techniques?

4. Methodology

4.1 Data pre-processing

First and foremost, raw data was extracted from the MITRE ATT&CK framework, specific to the groups and techniques in the Cloud Matrix. The dataset itself consisted of 86 groups and 53 techniques (sub-techniques included).

4.2 Data Mapping

For our procured data to be compatible with our FCA tools, the raw datasets were cleaned and processed using the R programming language, which lead to a formal context being exported as a .csv file for use within ConExp and .cex file in Lattice Miner. Lattice Miner and ConExp was utilised as our conceptual analysis tools with APT's/groups transposed as attributes and techniques as objects. Following the creation of the hierarchical structure of APT's and their applicable techniques it became apparent that mapping all 86 groups with applicable techniques would not be feasible within the scope of this particular research endeavour. In light of this, the approach to use only six of the most prominent APT's and the techniques associated with these six individual groups was deemed more effective. The approach enabled a more detailed view and underlying relationships could be showcased under a greater lens of focus.

4.3 Data Exploration

Viewing the relationships within Lattice Miner and ConExp exposed association and implication sets categorised and ranked according to their support and confidence levels. Low confidence is categorised as a percentage between 0.10 - 0.45. Medium confidence is seen as a percentage between 0.46 – 0.74 and high confidence, between 0.75 - 1.00. Antecedents and consequents of techniques and groups were found and noted. The level of support indicates the frequency of instances, and confidence the number of antecedents –consequent relationships occurring in each particular group and technique.

The Protégé ontology editor was used to substantiate these findings and showcase inferences within the data, in the form of an ontology. Under the main domain of the ontology concepts were initialised as follows: “APT”, “Proposition”, “Tactic” and “Technique”.

“APT” refers to the malicious entities implementing adversarial tactics. “Tactic” is the ultimate strategy which APT’s use to implement techniques. The six APT’s used are “Fox Kitten”, “Operation Wocao”, “Sandworm Team”, “APT32”, “APT41” and “Turla”. Each APT employs a technique through “techniqueEmployedBy”. “Technique” refers to the specific attack methods categorised under each tactic. Some techniques have sub -techniques indicated through “isSubTechniqueOf”. Each “Tactic” contains a certain number of techniques through “containsTechnique”. “Collection”, “CredentialAccess”, “DefenceEvasion”, “Discovery”, “Execution”, “Exfiltration”, “Impact”, “InitialAccess”, “LateralMovement”, “Persistence” and “PrivilegeEscalation”. “Proposition” represents the implementation which constitutes the relationship between the antecedent and consequent. “Proposition” contains six instances namely, “Adversary1”, “Adversary2”, “Adversary3”, “AR1”, “AR2” and “AR3”. Each proposition relationship includes the support and confidence level bound to it in the form of data property assertions, depicted as “hasSupport” and “hasConfidence”, followed by an assigned level ranging from “Low”, “Medium”, and “High”.

The given ontology, labels groups (APT’s) and techniques in a fashion which identically reflects its description on the MITRE ATT&CK framework. Further specification surrounding object properties include “asserts”, “hasProbability”, “implies”, “isAntecedentIn” and “isConsequentIn”. These object properties are associative entities defining linkages between the concepts/instances within classes.

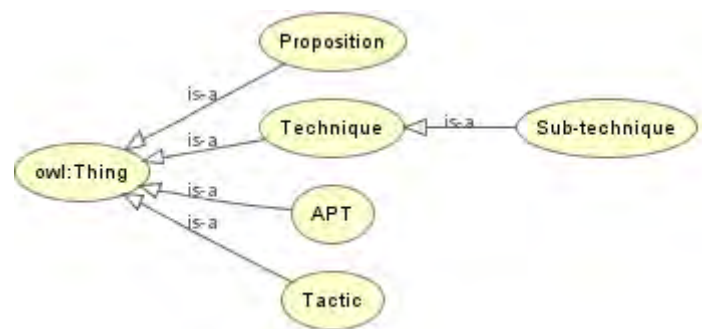


Figure 1: Diagram illustrating the Protégé OWL ontology

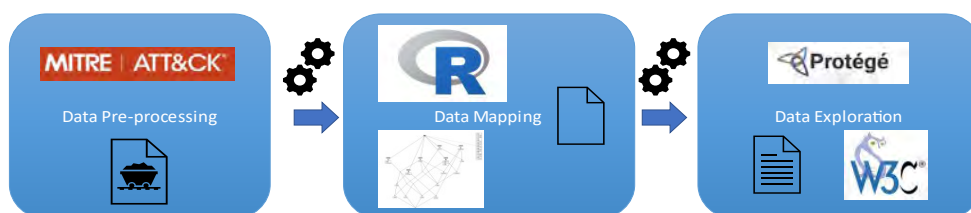


Figure 2: Diagram illustrating the research methodology

5. Data Analysis

5.1 Lattice Interpretation

Utilizing cyberattack techniques and APT groups from the MITRE ATT&CK Cloud matrix framework, a formal context, that focuses on the cyberthreats to cloud infrastructure and architecture was created. In this formal context the cyberattack techniques are objects, and the APT groups are attributes. In both ConExp and Lattice Miner the nodes of the lattice were labelled with names of the techniques used and the APT groups that used these techniques. In both figures (Figure 3 and Figure 4) a sample of six APT groups were chosen for an in-depth analysis of the formal context. In Figure 4 there are depicted concentrations of implications in the upper right region. The concentration of these implications indicates insightful connections on how APT groups and the techniques that they use relate to one another. From these connections, a more sophisticated comprehension

5.2 Implications and associations

It is imperative to evaluate the implications and associations between the APT groups in the lattices presented above. A relationship can be identified between the APT groups G0096(APT41), G0087(APT39), G0116(Operation Wacao) and G0117(Fox Kitten). The implication $G0096, G0117 \rightarrow G0087, G0116$ with three objects supporting the implication. In reference to the MITRE ATT&CK framework, the implication can be read as *APT41, Fox Kitten → APT39, Operation Wacao*. A deeper analysis indicates that all three groups made use of the technique T1190 (Exploit Public-Facing Application). This technique is implemented by APT groups exploiting vulnerabilities in internet-exposed computers or programs through the utilization software, commands or data in order to cause the application to operate not as intended. G0087 and G0117 are suspected to be based in Iran and G0096 and G0116 are suspected to be based in China.

In the utilization of implications provided by concept lattices, cyberattacks can be identified swiftly and efficiently as well as being able to determine the next steps or techniques that will be used in a cyberattack and aid in creating mitigations to prevent future cyberattacks.

The confidence level of association rules provides us with the probability that the implications will hold. The Lattice Miner solvers required a minimum support and minimum confidence value which were set as 0.0% and 100% respectively. Since there is a large number of associations, only a select number are highlighted. The different associations are grouped according to their confidence level percentage. The solvers show that the implication $G0096, G0117 \rightarrow G0116$ has a confidence level of 1.0 (100%), which indicates that there is a high probability that the techniques used by G0096 and G0117, will also be used by G0116. This illustrates that a greater focus would need to be placed on associations that have high probability values.

5.3 Ontology Queries and inferences

It is necessary to showcase the ontology utilization which provides the means for identifying APT's signatures or patterns in their cyberattacks on cloud infrastructure using knowledge representation. Example queries to illustrate the above mentioned will now be provided and discussed.

A security analyst suspects that a system has been attacked. The analyst determines the cyberattack technique used was T1078 (Valid Accounts). The analyst wants to determine which APT groups have utilized the same technique in order to identify the attacker. The analyst makes use of the ontology and writes the following query in order to achieve that:

employsTechnique some {T1078}

A security analyst wants to determine which tactic a technique falls under in order to better understand which phase an attack will undergo next. The query below indicates that the technique (T1078) is utilized in the tactics: initial access, persistence, privilege escalation and defence evasion.

containsTechnique some {T1078}

A security analyst can determine the APT group that utilizes techniques in implications with high confidence levels.

(employsTechnique some Technique) and isConsequentIn some (hasConfidence value "High")

A security analyst can determine which APT groups are likely to employ the same or similar techniques from querying implications and associations. AdversaryAR1 refers to the implication $G0096, G0117 \rightarrow G0116$.

isConsequentIn some {AdversaryAR1}

A security analyst is able to derive a list of cyberattack techniques from APT groups and from these techniques, a mitigation strategy can be designed and implemented.

techniqueEmployedBy some {G0096} or {G0117} or {G0116} and Technique

6. Results and Discussion

The MITRE ATT&CK Cloud Matrix has information on various cloud platforms such as Azure AD, Office 365, Google Workspace, SaaS, IaaS (MITRE ATT&CK, 2021), however the matrix does not contain all the information on cloud platforms as there are new cyberattacks daily with some going unnoticed and not recorded into the framework. By making use of Description Logic (DL) queries one is able to analyse relationships that go beyond just implications and associations between APT groups and cyberattack techniques. Taking into account support and confidence levels of associations, it enables for the identification of high priority cyberattacks. This is done through the utilisation of the ontology and formal context analysis results derived from the Cloud Matrix data. The concept lattice provides analysts with the necessary information on cyberattack technique concentrations in respect APT groups, allowing for analysts to better deal with the cyberattacks. The ontology is used to complement the concept lattice for better understanding and insight into how cyberattacks occur and proceed. Considering the fact that the ontology utilises open world reasoning as opposed to the FCA's closed reasoning, there is risk that unforeseen factors affect the ontological queries. Therefore, it should be noted that APT groups, techniques, implications and their associated support and confidence levels change with time. However, the usefulness of the concept lattice and ontology is still prevalent enough to make use of them in order to identify APT groups' signatures or patterns in their attacks.

7. Conclusion and Future Research

By supplementing the MITRE ATT&CK Cloud Matrix with FCA and implementing an ontological approach, the patterns and relationships between APTs and their techniques implemented could be analysed with greater understanding. This provides cybersecurity stakeholders a much more intuitive interaction with the ATT&CK Cloud Matrix, enabling faster response times and reduced impact as a result of cyberattacks.

The information and modus operandi depicted within this body of work contains foundational value as it could, without a doubt, be built upon in the sense to include more relevant and up to date information but also in the context of its' sophistication as an aid for cyber security practitioners. Further research should be conducted into expanding the body of research concerned with utilising AI knowledge representations in order to ascertain patterns or signatures of cyberattack techniques in order to identify APTs. By further refining this research and coupling it with better programmatic testing, new knowledge and insight can be achieved.

References

- Bertet, K., Borchmann, D., Cellier, P., & Ferre, S. (2007). Lattice Miner - A Formal Concept Analysis Tool. Rennes: Supplementary Proceedings of ICFC.
- Ganter, B., & Obiedkov, S. (2016). *Conceptual Exploration*. Berlin: Springer.
- Gronberg, M. (2019). *An Ontology for Cyber Threat Intelligence*. Ph.D. thesis, University of Oslo.
- Gruber, T. R. (2009). Ontology. In L. Liu, & M. T. Ozsu (Eds.), *Encyclopedia of Database Systems*. (pp. 1963-1965). Springer US.
- Kuznetsov, S., & Watson, B. (2017). Proceedings of the International Workshop on Formal Concept Analysis for Knowledge Discovery (FCA4KD). 1921. CEUR Workshop Proceedings.
- MITRE ATT&CK. (2021). *Cloud Matrix*. Retrieved June 8, 2021, from <https://attack.mitre.org/matrices/enterprise/cloud/>
- Mussen, M. (2015, June). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), 4-12.
- Obitko, M., Snasel, V., & Smid, J. (2004). Ontology Design with Formal Concept Analysis. 111-119.
- Pasquier, N. (2000). Mining association rules using formal concept analysis. 1867(12), pp. 259-264.
- Priss, U. (2006). Formal Concept Analysis in Information Science. *Annual Review of Information Science and Technology*, 40, 521-543.
- Red Hat. (2021). *What is cloud infrastructure?* Retrieved June 7, 2021, from <https://www.redhat.com/en/topics/cloud-computing/what-is-cloud-infrastructure>
- Singels, L., Biebuyck, C., & Malukele, L. (2020). A Formal Concept Analysis Driven Ontology for ICS Cyberthreats. (pp. 247-263). Hatfield: SACAIR 2020 Organising Committee.
- Strom, B., Applebaum, A., Miller, D., Nickles, K., Pennington, A., & Thomas, C. (2020, March). *MITRE ATT&CK: Design and Philosophy*. Retrieved June 2, 2021, from MITRE ATT&CK: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf
- Watson, L., & Watson, B. (2020). Exploratory data science on ATT&CK. *KM Conference 2021*. Leipzig: International Institute for Applied Knowledge Management.

Rising Above Misinformation and Deepfakes

Namosha Veerasamy and Heloise Pieterse

CSIR, Pretoria, South Africa

nveerasamy@csir.co.za

hpieterse@csir.co.za

Abstract: Misinformation can be rapidly spread in cyberspace. It thrives in the social media landscape as well as news platforms. Misinformation can readily gain momentum in the race to influence people or intentionally deceive. With the use of bots, misinformation can be easily shared, especially in environments like Twitter and Facebook. While, some measures are taken to stop the spread of misinformation, threats like Deepfakes are posing a higher challenge. Deepfakes provide a means to generate fake digital content in order to impersonate a person. With the use of audio, images and videos, artificial intelligence is used to depict the speech and actions of people. Deepfakes are typically made of presidents or influential businessmen such as Donald Trump and Mark Zuckerberg. Deep Fakes can be very realistic and convincing as this form of synthetic media is raising concerns about its possible misuse. The effects of Deepfakes are to spread disinformation, confuse users or create influence. This can lead to further effects like political factions, blackmail, harassment and extortion. Deepfakes could lead to a distrust in digital content as many may feel that anything we see is actually just a manipulation. Deepfakes has arisen as a new generation of misinformation through the manipulation of digital media in order to create realistic videos. This paper looks at the governing, communal and technical issues relating to Deepfakes. At the technical level, the use of audio and text analysis used to create Deepfake videos is advancing at a rapid pace which has also made its affordability and accessibility easier. An evaluation of the threats stemming from Deepfakes reveals that there are various mental, monetary and group dynamics involved. In this paper, the various types of threats emanating from Deepfakes is discussed. This paper also looks at five factors in the field of Deepfakes that should be taken into consideration (Technical Source Dissemination Victim Viewers). The paper discussed these five factors in order to help identify measures to help curb the spread of Deepfakes. A combination of these measures can help limit the spread of Deepfakes and support mitigation of the threat. Due to prominence and power that digital media has, it is imperative that this threat not be overlooked. The paper provides a holistic approach to understanding the risk and impact of Deepfakes, as well measures to help mitigate abuse thereof.

Keywords: Deepfakes, Misinformation, Voice Cloning, Text Synthesis, Machine Learning

1. Introduction

The manipulation of digital content is not a new phenomenon. However, improvements of machine learning technologies have given rise to a new generation of manipulated digital content – Deepfakes. The term “Deepfake” is an amalgamation of both “deep learning” and “fake” (Anon., 2021) and describes the production of highly realistic manipulated digital content. While most circulated Deepfakes are parodic in nature (Collins, 2019), a few occurrences have illustrated Deepfakes as an ideal platform for propaganda.

The proliferation of Deepfakes are driven by two factors: technological advances and societal context. Technological advances, such as increased computing power, 5G connectivity, 3D sensors, high-quality algorithms and pre-trained models, drive improvements in the quality of Deepfakes. From a societal perspective, the changing media landscape (a shift from the traditional model of centralised information distribution to user-generated content platforms) and the reliance on visual communication will cause Deepfakes to become a prominent source of information in the near future (Anon., 2021).

While the growing concern associated with Deepfakes remaining the distribution of misleading information, it is important to note that Deepfakes also provides potentially beneficial uses, especially for the entertainment industry. In *Gemini Man*, a 2019 hollywood blockbuster starring Will Smith, the special effects team reportedly relied on Deepfakes techniques to generate a younger but completely digital clone of Will Smith’s character (Bradshaw, 2019). Other positive uses of Deepfakes include voice-synthesis for medical purposes, animation of art to create virtual museums, as well as interactive educational lessons. However, the risks associated with Deepfakes necessitates continuous investigation of the technology.

The purpose of this paper is to investigate Deepfakes from the perspective of the following five factors: Technical, Source, Dissemination, Victim and Viewers. Such a broad understanding of Deepfakes is required to gather the necessary insight to identify applicable characteristics that can assist mitigation strategies. Mitigating the risks associated with Deepfakes, and manipulated digital content in general, requires continuous attention as Deepfake technology is a fast-moving target without any quick fixes currently available.

The remainder of this paper is structured as follows. Section 2 provides an overview of Deepfakes, focusing on the evolution of Deepfake technology, as well as highlighting notable Deepfake occurrences. In Section 3, the five factors of Deepfakes are discussed while Section 4 presents noteworthy characteristics of Deepfakes. Section 5 discusses mitigation factors that can be used to minimise the impact of Deepfakes as a form of misinformation and Section 6 concludes the paper.

2. Background

2.1 Evolution of Deepfakes

Commonly viewed by the general public as ultrarealistic fake videos (Harris, 2019), Deepfakes are better described as a subset of broader category of AI-generated “synthetic media”, including not only video and audio, but also photos and text (Anon., 2021). Deepfake technology first appeared on Reddit, a popular social media platform comprising of smaller communities used for content sharing and discussion. In November 2017, a user, known only as u/deepfakes, created such a community and shared the first rendition of the Deepfakes algorithm. Shortly thereafter, the community was renamed r/deepfakes, launching a new era of face-swapping videos (Fikse, 2018). The first collection of Deepfake videos contained pornographic content, in which the faces of the original actresses were replaced by well-known celebrities.

The first generation of Deepfake videos were characterised by: (i) low-quality synthesised faces, (ii) different colour contrast among the synthesised fake mask and the skin of the original face, (iii) visible boundaries of the fake mask, (iv) visible facial elements from the original video, (v) low pose variations, and (vi) strange artifacts among sequential frames. As of today, Deepfake videos have evolved both in quality and quantity due to improvements in technology and publicly available databases. The second generation of Deepfake videos considers different acquisition scenarios (e.g., indoors and outdoors), light conditions (e.g., day and night), distances from the person to the camera, as well as pose variations (Tolosana, et al., 2021).

Further advancements in Deepfake technology are demonstrated by the creation Deepfake videos using a single profile picture. Researchers at Samsung’s AI lab created a method to train a model using a single photo and various landmark facial features (e.g., shape of the face, eyes, mouth shape, etc.). While glitches were still clear and obvious, the technique elevates the risks of misinformation, deception and fraud to new levels (Solsman, 2019). Further research by Siarohin et al. improved on the technique by using a motion extractor that learns the extraction of key points along with their local affine transformations. Afterwards, a generator network models occlusions in the target motions and combines the appearance extracted from the source image and the motion derived from the driving video. (Siarohin, et al., 2019). The outcome is the animation of a static photo without requiring any prior information or knowledge of facial landmarks.

Technological improvements have enabled other forms of Deepfakes to evolve. Voice cloning technologies, also known as audio-graphic Deepfakes, speech synthesis or voice conversion/swapping (Neelima & Santiprabha, 2020), can generate sythetic speech that closely resembles a targeted human voice. Based on the concept of Text-to-Speech (TTS) technology, voice cloning relies on one of the following approaches: Concatenative TTS and Parametric TTS. Concatenative TTS uses a collection of audio recordings containing words and sounds to formulate sentences while Parametric TTS relies on statistical models of speech to simplify voice creation (Martin, 2021). As of today, advances in AI and deep learning have increased the quality of Parametric TTS-based voice cloning, leading to the creation of easily accessible and reusable AI-powered tools such as Tacotron, WaveNet, Deep Voice, and Voice Loop (Anon., 2021).

Text synthesis technologies relies on Natural Language Processing (NLP) to generate Deepfake text that imitates the unique writing and speaking style of a target. Key to NLP is a deep learning algorithm called the Transformer that allows for the transformation of input text into new text, by learning word sequences and sentence construction. One of the most advanced language models built on this architecture is Generative Pre-trained Transformer 3 (GPT-3), which was developed by OpenAI – an AI research laboratory. GPT-3 is a general-purpose NLP model consisting of 175 billion parameters that demonstrated notable performance related to translation, question-answering, as well as unscrambling words. However, researchers have found that the high-quality text generating capability of GPT-3 can make it difficult to distinguish synthetic text from the human-written text, leading to the potential misuse of language models (Brown, et al., 2020).

2.2 Notable Deepfake Occurrences

Over the past few years, multiple examples of Deepfake videos have appeared on YouTube, a popular video sharing and social media platform. Most Deepfake videos were created solely for entertainment purposes, however, a few key examples illustrate the potential power of the videos as a tool to spread misinformation.

One of the most prevalent and widely shared Deepfake videos stars Jordan Peele as former US President Barrack Obama. The video was created using FakeApp and involved taking an original video of Barack Obama and pasting Jordan Peele's mouth into it. The video, which took roughly 56 hours to develop, was created to raise awareness regarding the potential misuse of Deepfake videos (Silverman, 2018).

An example of a politically-motivated Deepfake video involved another former US President, Donald Trump, encouraging Belgium to withdraw from the Paris climate agreement. The Deepfake video was published by a Belgian social democratic party with the purpose of starting a public debat regarding the necessity to address climate change. The video was eventually debunked by Lead Stories (Lytvynenko, 2018).

In 2019, a video appeared on Instagram falsely portraying Facebook's Chief Executive Officer, Mark Zuckerberg, crediting a secretive organisation for the success of Facebook (Shead, 2020). The video was created by Canny AI, an Israeli company, and constructed using deepfake technology, an actor's voice and 2017 footage of Mark Zuckerberg (Metz & O'Sullivan, 2019).

The evolution of Deepfake technology was further demonstrated by criminals using AI-based software to impersonate the voice of a chief executive officer (CEO) and demand a fraudulent transfer of \$243000. The target, a CEO of a UK-based energy firm was convinced the caller was the CEO of the firm's German parent company and, therefore, approved the request. The criminals relied on the principles of authority and urgency to ensure the transfer took place within an hour (Stupp, 2019).

Finally, in 2021, various news outlets reported several European parliament members were targeted by Deepfake video calls imitating Russian opposition. The calls, which were created by Russian pranksters, Vovan and Lexus, were described as hyper-realistic and impressive (Roth, 2021).

3. Five Factors of Deepfake

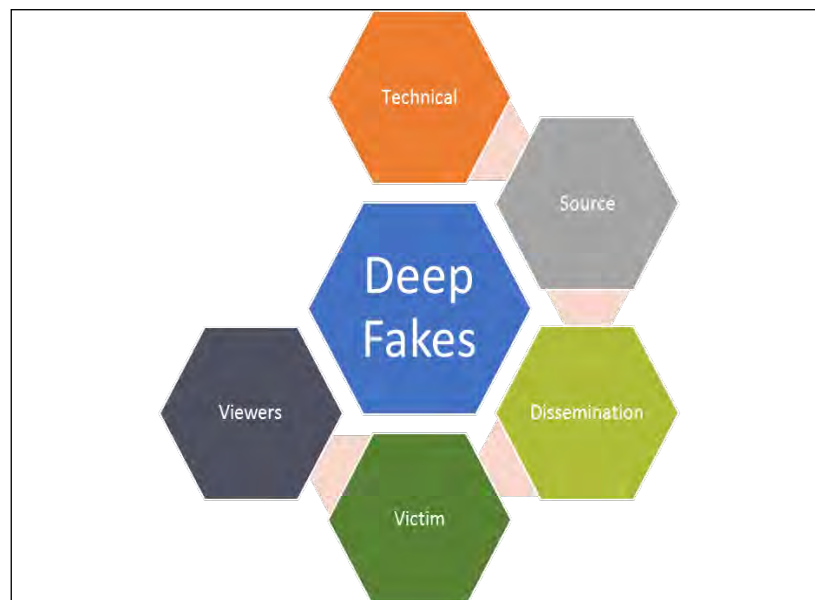


Figure 1: Five factors in field of Deep Fakes

This paper looks at five key factors in the field of Deepfakes that should be taken into consideration-see Figure 1 (Technical Source Dissemination Victim Viewers). The paper discussed these five factors in order to help identify measures to help curb the spread of Deepfakes. A combination of these measures can help limit the spread of Deepfakes and support mitigation of the threat. Due to prominence and power that digital media has, it is imperative that this threat not be overlooked. These five factors are discussed next

3.1 Technical

There are three technological factors contribute that contribute to the current Deepfake phenomenon: (i) AI concepts that leverages powerful machine learning (ML) and deep learning techniques, (ii) increased computing power, and (iii) datasets needed to train the algorithms.

3.1.1 Algorithms

Photo and video-graphic Deepfakes are commonly constructed using Artificial Neural Networks (ANNs), which are computing systems inspired by biological neural networks (Jain, et al., 1996). The basic processing elements of ANNs are a collection of artificial neurons, called nodes, which are connected to form edges and enable transmission between the neurons. Effects on edges are represented by weights (real numbers) and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. Both generative Adversarial Networks (GANs) and Autoencoders are examples of ANNs used to create photo and video-graphic Deepfakes (Abraham, 2005).

GANs were designed by Ian Goodfellow and his colleagues in 2014 (Goodfellow, et al., 2014) and are based on game theory involving two ANNs: generator and discriminator. A 'generator network' analyses a training set and identifies common patterns while a 'discriminator network' aims to determine whether the created content is similar by identifying forgeries. Detected forgeries are reported and the 'generator network' will attempt to improve the authenticity (Goodfellow, et al., 2020). The methodology provided by GANs have proven to be common use for the development of Deepfakes (Rana & Sung, 2020). Autoencoders represents another form of ANNs that are commonly used to construct Deepfakes. Any Autoencoder consists of two components: encoder and decoder. The encoder extracts the most important features to recreate the original input (e.g., photo of a person). To decode the features, separate decoders are generated for person A and person B, which are trained using a neural network. Once training is completed, face swapping can be performed by replacing the original decoder for person A with the decoder constructed for person B (Hui, 2018).

3.1.2 Computing Power

Successful development of Deepfakes requires appropriate computing power needed to execute the above-mentioned algorithms. For example, the supercomputer developed for OpenAI's GPT-3 is a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server (Langston, 2020). However, the current proliferation of Deepfakes illustrate the increasing availability and affordability of computing resources (e.g., CPUs, GPUs, or high-performance supercomputers). The cost associated with high-performance computing has reduced rapidly, driven by the increasing availability of cloud computing services (Collins, 2019).

3.1.3 Data Requirements

The final technical requirement for generating Deepfakes is an adequate training set. Traditionally, the development of Deepfakes relied on large volumes of data – the more training data available, the better the quality of the Deepfake. The Internet has contributed significantly to the generation of training sets with public figures (e.g., former US Presidents and celebrities) becoming the most favoured targets due to their prominent online digital footprint. However, recent research have demonstrated the possibility of generating Deepfakes using smaller datasets (Siarohin, et al., 2019, Zakharov, et al., 2019).

3.2 Source

The source of Deep Fakes can also be labelled the perpetrator. In many instances, the identity of the source is hidden and thus the source cannot be held accountable for actions taken. Deep Fake footage can be created by technology enthusiasts, researchers, academics, state-sponsored actors, political groups, lone actors, and even pornography developers. In terms of mitigation measures, legal frameworks may need to be developed in order to classify Deep Fakes as criminal acts. Also, diplomatic relations and international agreements may need to include members declaring that such technology will not be used for opportunistic or malicious purposes.

The creators/source of Deep Fakes mainly seek to achieve some of the following objectives:

- Misconstruction of truth
- Spread of fake news
- Mislead
- Creation of shock or alarm

- Discredit high profile individuals
- Entertainment
- Fraud
- Manipulation of events like elections
- Intimidation or blackmail
- Damage to stability or relations

The creators of Deep Fakes mainly seek to create impactful or destructive content that can have a consequential effect on individuals, companies and groups of people. Deep Fakes target individual persons but the effects can felt at a wide-scale level. The cascading resultss can be seen at different levels (see Fig 2).



Figure 2: Cascading Effects of Deep Fakes

3.3 Dissemination

Social Media (Facebook and Youube) as well as online platforms are the key mechanisms that are used in the dissemination of Deep Fakes. These dissemination channels play a key role in the scale and severity of impact. It is thus crucial that these online domains take some responsibility and carry some obligation in preventing the misuse of Deep Fakes. It is imperative that deepfake detective technologies be deployed, acting on the results as well as taking down identified malicious Deep Fakes. The circulation cycle needs to be adjusted in order to prevent rampant distribution and the negative effects thereafter. One of the most effective solutions may come from major tech platforms like Facebook, Google and Twitter voluntarily taking more rigorous action to limit the spread of harmful Deep Fakes (Toews, 2020). Efficient labelling and take down procedures should be in place.

Facebook enlisted researchers from Bekely, Oxford and other institutions to build a deepfake detector and help implement a ban as well as Twitter making policy changes to tag any Deep Fakes that are not removed (Adee 2020). In February 2020, Youtube declared that it would also be banning deepfake videos relating to the US election, voting process and US census (Adee 2020).

Furthermore two programs called Reality Defender and Deepttrace can be used to block Deep Fakes. Deepttrace uses an APT that will act like a hybrid antivirus/spam filter prescreening incoming media and diverting obvious manipulations to a quarantine zone, similar to how Google diverts spam (Adee 2020). Reality Defender comes from the company AI Foundation and aims to tag and contain manipulated images and video before they cause damage.

3.4 Victim

A core target of Deep Fakes are women. Many victims of image-based sexual abuse—also known as revenge porn—or deepfakes feel powerless (Fighthenewdrug.org, 2021). Nonconsensual pornography accounts for 96% of deepfakes currently on the Internet (Adee, 2020).

However, women are not the only target. Deep Fakes can be used to bully, intimidate, trick or mock victims.

Many companies may be concerned about Deep Fakes due to the potential use for fraud and theft. Deep Fakes can be used to convince people in power to make payments to fraudsters. Extortion could be another key area of concern. Deepfakes could be used to make fraudulent payments or hack into individuals banking services.

At a national level, Deep Fakes could threaten democracy. Shocking video footage could be created of electoral candidates in order to smear their campaigns. Celebrities are also targeted and Deep Fakes discredit their reputation. In many cases, Deep Fakes are created of prominent figures in order to create misinformation, manipulate the public, mislead or generate shock or confusion.

From the perspective of protection, establishment of support mechanisms for victims is needed, as well as improving the capacity of data protection authorities together with encouraging the protection of fundamental human rights.

3.5 Viewers

Deep Fakes have a cascading impact from the individual to group dynamics, as well as at societal and community level. Deepfakes threaten to grow from an Internet oddity to a widely destructive political and social force and society needs to act now to prepare itself (Toews, 2020). Viewers of the Deep Fakes may react in different ways. Some may choose to believe the footage and lose trust in the individual or company. A deepfake can leave an imprint and hurt someone's reputation especially if their name and face is part of a negative video or audio-real or deepfake (Somers 2020). In terms of protection, it is important that investment is made in labelling trustworthy sources and media and security awareness so that the public is able to identify potential Deep Fakes.

4. Characteristics of Deepfakes

Carefully made deepfakes can already be very realistic, though only under certain circumstances—an attentive observer will notice that convincing deepfakes focus on individuals who don't wear glasses or have beards, and typically use a stationary camera (Engler, 2019).

In some instances by studying videos carefully, one may be able to detect that the footage is a Deep Fake. Some characteristics to look for (Kaspersky, 2021)

- Jerkey movements
- Shifts in lighting from one frame to the next
- Shifts in skin tone
- Strange blinking or no blinking (or long pauses between blinking)
- Poorly synched lips and sound
- Digital artifacts in the image

Sample mentions some of the following signs to look out for as well (2020):

- Bad lip synching
- Patchy skin tone
- Flickering around edges of transposed faces
- Fine details like hair are difficult to render well
- Badly rendered jewelry or teeth
- Strange lighting effects
- Inconsistent illumination and reflections on the iris

5. Mitigation Factors

In the security world, there is continuous dynamic of the creation of protection mechanisms which attackers then try to figure out a means to bypass. With Deep Fakes, we find the same paradigm. For example, a filter can be added over an image so that the image cannot be used to create a deepfake. However, innovative technologists may find ways to circumvent controls and still accomplish their goal of Deep Fake generation.

Another protection mechanism is digitally signing content. Similar to a security certificate, content can be validated through a third party. The challenge here is that in some cases, viewers will not be interested in whether the content is digitally signed. For example, revenge porn.

Another emerging technology is the use of cryptographic algorithms to insert hashes at certain intervals in the video; if the video is modified the hash will change (Kaspersky, 2021). This is comparative to watermarking documents.

Another preventative technique is to use a program to insert specially designed digital artifacts into the videos to conceal the patterns of the pixels that face detection software uses (Kaspersky, 2021). This will slow down Deepfake algorithms or lead to poor quality videos.

The International Risk and Governance Center has proposed several measures to help govern Deepfake risks. These include (Collins, 2019):

- Detection: Continued research into technologies to distinguish between authentic and fabricated digital content
- Provenance: Techniques designed to verify the origin and integrity of digital artefacts, such as trusted-hardware schemes or ways of preserving metadata
- Image rights and control: Greater control for individuals over digital content that relates to them, including potential “takedown” rights
- Digital corroboration: The use of multiple independent data sources, analogous to the familiar process of corroborating eye-witness testimony
- Secure digital processes: A greater focus on authentication and verification to make digital communication less vulnerable to deepfakes
- Platform nudges: Interventions to influence the way people — and algorithms — share digital content
- Legal guidance- clarification of the ways in which existing legal frameworks apply to deepfakes
- Penalties- The persistence nature of some harms involving digital content may require changes in the way they are penalised

Technology platforms have a key role to play in the spread of Deep Fakes. Efficient labelling and take down procedures should be implemented. A platform can flag a Deep Fake once detected. This can help reduce the spread of harmful content.

At a policy level, measures should be put in place like extension of legal frameworks to include Deep Fakes as criminal offenses, as well as international agreements to refrain from the use of Deep Fakes. Furthermore, more stringent measures like the imposition of sanctions on states involved in disinformation and Deep Fakes could help curb the spread.

Furthermore at a research level, tremendous investment is needed to develop AI techniques that can block, slow or obscure Deep Fake creation.

Overall, awareness will also be needed in order to educate the public about the possibilities and dangers of Deep Fakes. An informed community may form a crucial aspect of defending against misinformation from Deep Fakes.

6. Conclusion

It is expected that more Deep Fakes will be distributed in order to manipulate public opinion, carry out distortion, manipulation of elections, undermining journalism, inflicting reputational damage, destabilise and cause misinformation. The growth of Deep Fakes may erode public faith in digital media. The effects of Deep Fakes will be spread of disinformation, increase in doubt as well as ability to deny accountability (claiming material is fake). The public may find it increasingly difficult to differentiate from what is real and what has been maliciously generated. We live in a world undergoing a rise in disinformation and fake news. The modern version of photoshopping, users can now impersonate a politician, appear in a movie or dance like a professional. Deep Fakes may grow rapidly. There are no easy solutions. Overall, more continuous efforts and governance is needed in order to curb the prominent spread of Deep Fakes. Technology providers need to take a more active role in the proper use of their technologies. Policy makers and regulators need to implement stronger measures in order to protect victims and punish perpetrators. Legal frameworks need to include the criminal implications of malicious Deep Fakes. International agreements should concur that members will refrain from the use of Deep Fakes. Furthermore, an additional consideration is the imposition of sanctions on states engaged in disinformation and Deep Fakes. Visual manipulation has emerging as a fast-moving trend. It is imperative that authorities and specialists around the world put in a concerted effort to mitigate its rampant growth.

References

- Abraham, A., 2005. Artificial neural networks. In: P. H. Sydenham & R. Thorn, eds. *Handbook of measuring system design*. s.l.: John Wiley & Sons, pp. 901-908.
- Adee S, April 29 2020, What are Deepfakes and how are they created?, IEEE, Available at <https://spectrum.ieee.org/what-is-deepfake>, [Accessed 27 September 2021].
- Anon., 2021. *Tackling deepfakes in European policy*, s.l.: European Parliamentary Research Service.
- Brown, T. et al., 2020. *Language models are few-shot learners*. s.l.: arXiv preprint arXiv:2005.14165.
- Collins, A., 2019. *Forged Authenticity: Governing Deepfake Risks*, Lausanne: EPFL International Risk Governance Center.
- Engler A, November 14 2019, Fighting deepfakes when detection fails, Available at: <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/>, [Accessed 27 September 2021].
- Fightthenewdrug.orgviewer, 2021, 7 Things you can do if you're a victim of Deepfakes or revenge porn, Available at : <https://fightthenewdrug.org/7-things-you-can-do-if-youre-a-victim-of-deepfakes-or-revenge-porn/>, [11 October 2021].
- Fikse, T. D., 2018. *Imagining Deceptive Deepfakes*. Oslo: University of Oslo.
- Goodfellow, I. et al., 2014. Generative adversarial nets. *Advances in neural information processing systems*, Volume 27, pp. 1-9.
- Goodfellow, I. et al., 2020. Generative adversarial networks. *Communications of the ACM*, 63(11), pp. 139-144.
- Harris, D., 2019. Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review*, Volume 17, pp. 99-127.
- Hui, J., 2018. *How deep learning fakes videos (Deepfake) and how to detect it?*. [Online] Available at: <https://jonathan-hui.medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9> [Accessed 20 September 2021].
- Collins, Aengus. (2019). *Forged Authenticity: Governing Deepfake Risks*. Lausanne: EPFL International Risk Governance Center.
- Jain, A., Mao, J. & Mohiuddin, K., 1996. Artificial neural networks: A tutorial. *Computer*, 29(3), pp. 31-44.
- Kietzmann, J., Lee, L. W., McCarthy, I. P. & Kietzmann, T. C., 2020. Deepfakes: Trick or treat?. *Business Horizons*, Volume 63, pp. 135-146.
- Kaspersky, 2021, Deepfake and Fake Videos- How to Protect Yourself, Available at: <https://www.kaspersky.co.za/resource-center/threats/protect-yourself-from-deep-fake>, [Accessed 27 September 2021].

- Langston, J., 2020. *Microsoft announces new supercomputer, lays out vision for future AI work*. [Online]
Available at: <https://blogs.microsoft.com/ai/openai-azure-supercomputer/>
[Accessed 20 September 2021].
- Lytyvnenko, J., 2018. *A Belgian Political Party Is Circulating A Trump Deepfake Video*. [Online]
Available at: <https://www.buzzfeednews.com/article/janelytyvnenko/a-belgian-political-party-just-published-a-deepfake-video>
[Accessed 18 September 2021].
- Martin, K., 2021. *What is Voice Cloning*. [Online]
Available at: <https://www.idrnd.ai/what-is-voice-cloning/>
[Accessed 20 September 2021].
- Metz, R. & O'Sullivan, D., 2019. *A deepfake video of Mark Zuckerberg presents a new challenge for Facebook*. [Online]
Available at: <https://edition.cnn.com/2019/06/11/tech/zuckerberg-deepfake/index.html>
[Accessed 18 September 2021].
- Neelima, M. & Santiprabha, I., 2020. *Mimicry Voice Detection using Convolutional Neural Networks*. s.l., s.n., pp. 314-318.
- Rana, M. & Sung, A., 2020. *Deepfakestack: A deep ensemble-based learning technique for deepfake detection*. s.l., IEEE, pp. 70-75.
- Roth, A., 2021. *European MPs targeted by deepfake video calls imitating Russian opposition*. [Online]
Available at: <https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition>
[Accessed 18 September 2021].
- Shead, S., 2020. *Facebook to ban 'deepfakes'*. [Online]
Available at: <https://www.bbc.com/news/technology-51018758>
[Accessed 18 September 2021].
- Siarohin, A. et al., 2019. First Order Motion Model for Image Animation. In: *Advances in Neural Information Processing Systems*. s.l.:Curran Associates, Inc., pp. 1-11.
- Silverman, C., 2018. *How To Spot A Deepfake Like The Barack Obama-Jordan Peele Video*. [Online]
Available at: <https://www.buzzfeed.com/craigsilverman/obama-jordan-peelee-deepfake-video-debunk-buzzfeed>
[Accessed 17 September 2021].
- Solsman, J., 2019. *Samsung deepfake AI could fabricate a video of you from a single profile pic*. [Online]
Available at: <https://www.cnet.com/tech/computing/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/>
[Accessed 13 September 2021].
- Somers M, July 21, 2020, Deepfakes Explained, MIT Management Sloan School, Available at :
<https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained>, [Accessed 11 October 2021].
- Stupp, C., 2019. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*. [Online]
Available at: <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>
[Accessed 18 September 2021].
- Toews R, May 25 2020, Deepfakes are going to wreak havoc on Society. We are not prepared, Forbes, Available at:
<https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=2f9027c17494>, [Accessed 27 September 2021].
- Tolosana, R., Romero-Tapiador, S., Fierrez, J. & Vera-Rodriguez, R., 2021. *Deepfakes evolution: Analysis of facial regions and fake detection performance*. s.l., Springer, pp. 442-456.
- Zakharov, E., Shysheya, A. & Burkov, E., 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *CoRR*.

Multi-Purpose Cyber Environment for Maritime Sector

Gabor Visky¹, Arturs Lavrenovs¹, Erwin Orye¹, Dan Heering², Kimberly Tam³ and Olaf M. Maennel⁴

¹NATO Cooperative Cyber Defence Centre of Excellence, Tallinn, Estonia

²Estonian Maritime Academy, Tallinn University of Technology, Estonia

³University of Plymouth, UK

⁴Department of Computer Science, Tallinn University of Technology, Estonia

gabor.visky@ccdcoe.org

arturs.lavrenovs@ccdcoe.org

erwin.orye@ccdcoe.org

dan.heering@taltech.ee

kimberly.tam@plymouth.ac.uk

Abstract: The cyber attack surface in a maritime environment is constantly growing. More current information and computer technologies are being used on cargo and passenger ships to save on operational costs and increase navigational safety. Along with the growing reliance on automation, the risk of a disruption to a vessel's critical systems by drawing on the wrong inputs from sensors to change the behaviour of the actuators has significantly increased. Traditional operational technological systems are much more complicated to update than the automatic software updates we see in information technology systems. To better understand existing cyber threats in the maritime sector and increase cybersecurity resilience, this paper aims to replicate the digital components of a ship's bridge to examine scenarios when the bridge system loses connectivity, receives the wrong inputs from sensors, or the internal system becomes compromised. The simulator differentiates fundamentally from traditional simulators or digital twins in the maritime sector that focus on training seafarers. This environment generates data streams that are similar to those on board a ship. Those data streams can be analysed, modified and spoofed to observe the effects. The effects can be technical but it is equally necessary to analyse how human beings would react in specific circumstances. Our work provides the opportunity to isolate the ship network traffic, conduct penetration testing, find cybersecurity vulnerabilities on devices, and execute cyber attacks without the dangers associated with running such scenarios on a vessel in the open sea.

Keywords: maritime, cybersecurity, testbed

1. Introduction

Cyber threats and actual incidents in the maritime sector are constantly growing. Recent cyber incidents and accidents have highlighted how fragile the naval industry is, despite its importance in world trade. The industry depends increasingly on digitalisation, integration of systems, operations, and automation. The growing role of autonomous vehicles makes the problem more severe. The consequences of possible cyber attacks can include financial losses, safety issues, bad publicity, and compliance risks. The cases usually originate in technology, staffing, or cybersecurity operating procedures that lead to technology testing and education.

This paper proposes a multi-functional environment for cyber-related education and maritime-related cyber research that is flexible enough to adapt to specific needs. Although the proposed environment has enormous potential in education, this paper highlights the technical perspective and introduces the testbed functionality. The simulator aims not to provide sailing-related experience but focuses on the consequences of cyber attacks and how to react to those attacks. Furthermore, the environment offers a maritime-related climate for cyber experts to conduct experiments.

The environment has a network similar those found on vessels. The generated data streams traffic is used as a source for analysis and penetration testing. Using the simulator, we achieve similar results to those on actual vessels without violating the integrity of the expensive equipment on board a ship. Even though the proposed environment has considerable potential for education, this paper highlights the technical perspective and introduces the testbed functionality.

The first section focuses on the current situation in the maritime sector and its challenges. In Section 2, we provide relevant background on cybersecurity in the maritime sector. Section 3 reviews related research addressing the available solutions that can be used as a training or research environment. In Section 4, we describe the setup of the multi-purpose cyber environment that was built on the premises of the NATO

Cooperative Cyber Defence Centre of Excellence, Tallinn, Estonia, and discuss its similarities and differences compared to a real ship. Section 5 presents the conclusions and highlights the future directions.

2. Background

In recent times, the importance of cyber defence has been continuously increasing in the maritime sector, as in other industries. The following section provides an overview of the sector's specific characteristics.

2.1 Cybersecurity in the maritime sector

Nowadays, the importance of the shipping industry for modern society is constantly growing. The volumes of goods carried by ships in 2019 reached 11.08 billion tons (*Explaining shipping*, 2021). It is estimated that over 80% of world trade is carried by the shipping industry (United Nations, 2021). The need to reduce operational costs has led the shipping industry to seek new and advanced technological solutions. In 2017, the International Maritime Organization (IMO) initiated a scoping exercise to determine how the safety, security and environmental soundness of Maritime Autonomous Surface Ships (MASS) operations might be addressed using IMO instruments (International Maritime Organization, 2021). The development of technologies and policies that make autonomy a feasible solution has enabled several organisations to make progress on such projects and will soon sail their first autonomous ships (Tam and Jones, 2018).

Along with the growing reliance on automation, the risk of external interference and the disruption of critical systems is greatly increased; malicious actors can interfere with the different control systems of the ship. They can cut off all external communications or obtain confidential data. Cybersecurity on board ships is gaining in importance due to recent incidents on ships at sea (Caprolu *et al.*, 2020).

In June 2017, the world's largest container shipping company, A.P. Møller-Maersk was one of the companies hit by the malware NotPetya (WIRED, 2018). In July 2018, one of the biggest shipping companies, China Ocean Shipping Company (COSCO), was victim of a cyber attack (Goud, 2018). Norsk Hydro was hit by an extensive cyber attack in March 2019 (SAFETY4SEA, 2019). Although these incidents were ransomware related and the role of the operators was high, it has been successfully demonstrated by researchers that both information technology (IT) and operational technology (OT) systems used on board have vulnerabilities that can be exploited or exposed unintentionally by crew members (Bhatti and Humphreys, 2017), (Pen Test Partners, 2020).

Those unique circumstances demanded creating an environment for research and education in this field so that the maritime sector will become more resilient to cyber attacks.

2.2 Aspects of cybersecurity

Cybersecurity refers to the application of technologies, processes and controls to protect systems, networks, programs, devices and data from cyber attacks. In Information Technology (IT) it focuses on the practice of ensuring the confidentiality, integrity and availability of information (CIA triad). It comprises an evolving set of tools, risk management approaches, technologies, training, and best practices designed to protect networks, devices, programs, and data from attacks or unauthorised access (Gourlay, 2000).

Cybersecurity has several aspects: application security, information or data security, network security, disaster recovery/business continuity planning, operational security, cloud security, critical infrastructure security, physical security, and end-user education. Technology, people, and regulatory frameworks are always essential to solving problems in cybersecurity, and this is no different in the maritime field. These three elements are fundamentally interconnected. At the same time, research can improve how these function from a cybersecurity perspective for the maritime sector.

According to (Fortress Information Security, 2020), the maritime-specific cybersecurity challenges span technology, staffing, and cybersecurity operating procedures. The main purpose of this work is to provide an environment that can be used to educate the crew, research technologies and test operating procedures.

2.3 Technology-related challenges

Maritime organisations have a vast array of legacy operational technology (OT) and information technology (IT) systems deployed. The majority of those systems were not designed with cybersecurity in mind. In addition, the different OT system configurations and architectures on the various ships make it challenging to secure the network infrastructure and their topologies aboard a vessel.

The situation is becoming more severe since autonomous or remotely operated off-shore vehicles will contain even more OT components, interconnected with navigational and other subsystems. Since the control system of the whole industry cannot change overnight, new ships have to maintain compatibility with obsolete systems, and since they have the heritage of old solutions requested by legacy standards, the applicability of the newest cybersecurity solutions will be limited.

Despite growing automation on ships, communication links use satellite-based solutions that provide only limited bandwidth. These circumstances limit remote assistance and maintenance in the case of a cybersecurity problem.

Appropriate technologies (e.g. antivirus, firewalls, intrusion detection/prevention systems, endpoint security and others) should be selected and implemented to provide comprehensive cybersecurity protection for these OT systems. Still, ship-system producers keep sailing safety in the primary position, often at the cost of cybersecurity. In most cases, it is hard to select the proper defence solution for several reasons.

The applied technology is barely standardised, the communication standards are vendor-specific, and furthermore, with an approximated 40-year life cycle, operators on ships cannot update CPS on the same frequency we are doing with generic IT systems where software vendors are constantly pushing their updates. Therefore, it is difficult to neutralise potential vulnerabilities in this field and this leads to weak system resilience. Currently, there is a lack of environments in which we can research and educate the maritime sector to test the different components individually or the whole system.

2.4 Education-related challenges

Ships, ports, terminals and offshore facilities are increasingly dependent on networked information and communication technology (ICT) (Heering *et al.*, 2020). Seafarers must be ready to cope with a growing number of cyber threats on board ships, with cybersecurity awareness playing an essential role in emergency and crisis management. Maritime education and training institutions (MET) offer high-level facilities for training seafarers, such as maritime simulators. These tools aim at learning and practising different operations like sailing, docking, and the use of onboard devices like RADAR, Electronic Chart Display and Information System (ECDIS), and others. Cybersecurity education is usually outside the scope of the curriculum, and the facilities do not support this task.

The authors argue that a certified ship bridge simulator can be used for teaching and practising responses to cyber-related incidents. Besides this fact, the lack of a proper testing environment is a severe barrier to improving cybersecurity awareness. The key to more cyber-safe operations at sea lies in the proper implementation of cybersecurity awareness training for active and future seafarers by taking into account their responsibilities onboard the ship and their background knowledge of IT. Training and awareness are key elements to effective cyber risk management on ships.

3. Related work

3.1 Testbeds

Determining the vulnerabilities of industrial control systems using embedded devices is a complicated process because of the complex hardware and software interactions. Lund *et al.* (Lund *et al.*, 2018) shed light on the great importance of the central components like Integrated Navigation Systems (INS) and ECDIS for the safety and security of maritime operations, but claims that the topic has barely been studied.

One approach is to build a comparatively simple system that captures the relevant complexity (i.e., a testbed) (Davis *et al.*, 2006). A testbed is an essential tool, avoiding the need to experiment exclusively on live systems. The topic is highly researched in the field software testing, primarily in the context of vulnerability identification. Salunkhe *et al.* have published (Salunkhe *et al.*, 2018) systematic literature review results regarding the cyber-physical testbeds focused on simulation. According to their findings, the literature focuses on electrical grids, network and communication, but not on maritime related topics.

Frank *et al.* introduce the design considerations for cybersecurity testbeds (Frank *et al.*, 2017) for education. The main requirements were automated deployability, reusability, low cost, high availability and scalability. The paper introduces the main challenges of the testbed setup and installation, but the cases focus on the vulnerabilities of web application, which is not relevant on a ship.

3.2 Maritime-related solutions

Brinkmann and Hahn offer a testbed architecture for maritime systems (Brinkmann and Hahn, 2017). Their work introduces the physical testbed 'LABSKAUS' (Laboratory for Safety Critical Analysis On Sea) as part of the eMaritime Reference Platform (eMIR), which is a CPS, and provides various maritime specific components, such as a reference waterway, research boat, sensor infrastructure and a mobile bridge. According to the paper the development of safety-critical systems such as a highly automated and autonomous vessel brings the need to establish a test environment (or 'testbed') close to the real world in addition to simulative test environments. This solution can support the development, validation and certification process. The highly sophisticated testbed uses a sensorbox that contains sensors and actuators which correlate with the environment; for example, a waterway network or a hydrodynamic environment. The complex system supports the solution of technical questions, but does not have educational features.

Modelling and simulation technologies are also extensively used in the maritime industry. A digital twin is a digital representation of a physical object, asset or system: a ship, a car, a wind turbine, a power grid, a pipeline, or a piece of equipment, such as a thruster or an engine (Smogeli, 2017). Although they enable early and continuous simulation-based testing, cutting the expenses of the system integration (Smogeli, 2017), this solution cannot be used to identify vulnerabilities, since the digital twin setup differs from the actual devices installed on ships. At the same time, a system based on a digital twin can be used for cybersecurity education.

Tam et al. presents a Cyber-SHIP Lab (Hardware, Software, Information and Protections) as a next-generation research capability for maritime cybersecurity (Tam *et al.*, 2019). This facility offers a complex research capability considering the physical aspects as well as the digital, with a lab that is accurate at the hardware level and not based on simulation or emulation. According to the plans, Cyber-SHIP would accommodate ship controller devices, so ship-identical systems could be installed at a considerably higher budget. Multiple configurations of the equipment can be created to imitate different ships. Although the setup proposed in our paper has the same objectives (supporting research and education), but our solution uses simulation-based data source that offers a more simple setup but still close to the real world.

Although the relevant literature is rich, only a few papers are available about a maritime-related environment for education, research and penetration testing. Their research offers help for industry, government, and academia to understand and mitigate cyber threats in the maritime sector.

4. Multi-purpose cyber environment

As was introduced in the previous chapter, the maritime industry faces challenges in cybersecurity. These issues are also growing instead of getting easier to solve. The following section presents the multi-purpose cyber environment, a tool that can be used to ease the pain related to those problems.

The solution can be used for several different purposes. Besides the demonstration and educational functions, it offers data source functionality and environment vulnerability testing.

4.1 System setup

The environment is based on a Transas NTPRO 5000 Navigational Simulator that was designed exclusively for educating crew. A simulator environment contains the visualisation part that is responsible for the audio and visual experience, with several TVs, as it can be seen on Figure 1.

All the environmental parameters that can influence the simulated ship, such as visibility, wind and weather conditions, current and tide, can be set on the instructor machine, just like marine traffic and the position of other ships, that defines the RADAR picture and NMEA-Messages from real world data (e.g. Automatic Identification System (AIS)).



Figure 1: Multi Purpose Cyber Environment

The seafarer, who operates the simulated ship can control it via its fundamental components – the wheel, buttons and telegraph – and can see the actual status of the vessel on the screen of the coning machine. All the parameters are calculated by the simulator server. The crew can use the MFD as on a real ship. Since the laboratory is used for demonstration purposes as well, an extra monitor is added on which the screen of the MFD can be seen. As Figure 2. shows, the original setup was extended using a Gateway Machine, Researcher's Computer, Data Collection Unit and additional MFD for research purposes.

In a simulator environment, the central components (i.e. MFD and ECDIS) are identical to the devices installed on ships, but the computers contain simulator-related software components as well. These software components (RADAR and MFD Server) generate a ship-like network traffic according to the simulated values; therefore, the network traffic of the sensors and RADARs exists only in these computers and are not available from outside of the device.

In contrast to a real ship, the sensors measure the different physical values that are converted into a digital format and transferred via NMEA-0183 protocol to the sensor integration unit – though given different names such as Data Distribution Unit, Data Acquisition Unit, Data Collection Unit (DCU), Sensor Concentration Unit and so on – that converts these values into an IP based data communication format, usually TCP. (In some cases, the sensors have direct serial connections to the workstations) Since the RADAR data (picture) is different from the sensor data, the RADAR is treated differently: it is connected directly to the INS. RADAR can have an independent network or it can use the same network but a different protocol, for example UDP.

The main purpose of this research was to create an environment, where the ship network traffic appears, and the control devices are exactly the same (without simulator-related software pieces) like on a vessel. This solution provides the opportunity to isolate the ship-network traffic, conduct penetration testing on devices, or simulate cyber attacks. To achieve this goal, a Gateway Machine is used to host the RADAR and MFD servers, emulate the ship's RADAR and the DCU from which the ship-related data is available. The simulator sends the simulated data to the Gateway Machine, and the MFD and RADAR application also connects to it and reads the sensor values and RADAR pictures.

Since the simulator offers a wide range of ship types, the solution provides configurable network traffic that is very close to the actual ship's network traffic.

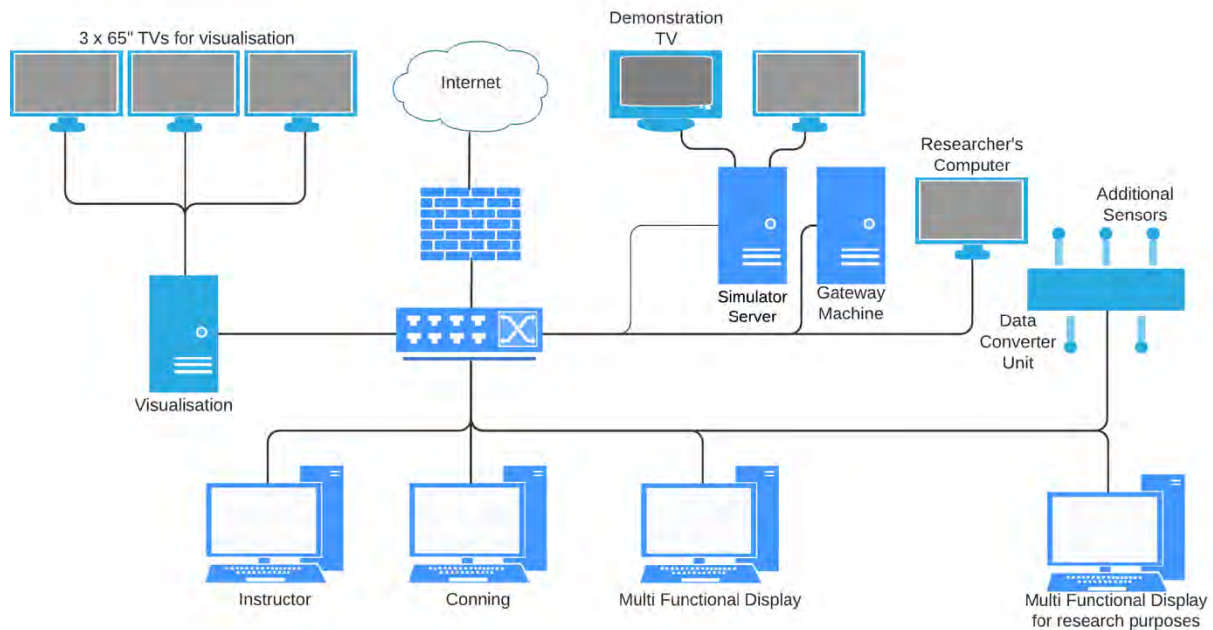


Figure 2. System setup.

4.2 Use cases

The multi-purpose cyber environment can be used mainly in the field of technology and education. This paper focuses on technology-related use cases that will be introduced in detail in this chapter.

4.2.1 Vulnerability testing environment

Nowadays, a ship can have several isolated networks for different purposes. The navigational network, the ship control network, the WiFi network used during loading and unloading the cargo, and the propulsion controlling network are all isolated networks (sometimes physically). The networks are logically interconnected by different crew members, who operate different devices based on data from different sources. As the level of integration between ships is increasing, these networks could no longer be separated. With less system isolation to protect systems, environments for simulating complex networks become more useful to the maritime sector.

The obsolete, twisted-pair and serial data transmission based on IEC 61162-1/NMEA-0183, and more recently the NMEA-2000 protocols are widely used on vessels. Since the devices that use these protocols must be integrated on modern ships, the legacy protocol was encapsulated into TCP/IP packets. Although this solution was simple, it brought a considerable amount of cyber-related problems into the picture. The modern ships using TCP/IP networks use legacy NMEA-0183 protocol packets encapsulated into TCP packets to transport their sensor data. The multi-functional environment can be used for generating this kind of network traffic, which can be analysed and reused in different ways.

The highly sophisticated switch offers Encapsulated Remote SPAN (ERSPAN) capability. As the name says, it delivers generic routing encapsulation (GRE) for all captured traffic and allows it to be extended across Layer 3 domains. This feature together with the installed VNC clients enables full remote and network traffic analysis via a Virtual Private Network (VPN).

The Transas NTPRO-5000 – on which the environment is based – offers several different ships with its sensor and actuator sets for selection and the available sensor data generated according to the selected ship's sensor set. In the ship's network segment a research computer is installed for network traffic analysis, creation, and modification purposes. This solution offers an environment, where the ship-identical network data is available for the tested device, together with a modified or artificially generated one.

The environment also helps develop and test cyber defence solutions, such as antivirus, intrusion detection systems, and so on, since all the necessary devices are available for measurement and behaviour analysis in the setup.

4.2.2 Offensive cyber testing environment

There exist dozens of different types of cyber attacks, and defending against them is a tedious task with no single overall solution available. Since ships can be considered a system of systems, their defence, especially via the IT infrastructure and port management systems, is demanding. A ship's controlling device is an industrial control system (ICS). These are specialist information systems that differ significantly from traditional information systems used in the IT world (Drias *et al.*, 2015). Some attacks are not applicable against a ship; for example, phishing attacks, URL interpretation attacks, or web attacks.

In a ship, one primary defence against cyber attacks is air-gapping ship controls, navigation, and sensors from the rest of the IT systems, networks and the internet. Correctly implemented air-gapping creates a challenge for the attacker. If there is no possibility of a network connection, the remaining way to compromise the system is to deliver malware via a physical medium such as a thumbdrive, which is connected to and physically transferred between the air-gapped system and other ship or off-ship systems that are compromised and ready to deliver the malicious payload.

A scenario with an inside attacker significantly expands the attack surface on the air-gapped system. An insider might purposefully craft and deliver malware via physical media, or even do it off-ship or instruct somebody else to deliver it. An insider might break the air-gapping by physically connecting to another ship network that can have deployed malware on other connected computers making it less noticeable. Air-gapping could be broken permanently, temporarily (e.g., to deliver malware), or intermittently – installing a communication device that could receive commands from and send data to an off-ship location. In the simplest case, it could be a mobile router or modem with a mobile data roaming plan that works across the globe and can communicate in ports or near the shore, or any of the vast variety of point-to-point communication devices enabling communication in more fixed locations or between moving ones (e.g., from another ship).

In the simplest case, these communication devices can be connected directly to one of the computers in the air-gapped network and be used by already functioning malware. In this case, the communication device only provides a communication channel for the malware, the computer provides power and the network stack. The malware injects network packets or manipulates the system's inputs or outputs. This type of deployment is the most visible at the system level – it might be visually displayed, visible in the systems network configuration and the malware might also be detected by defences on the infected system.

Other cases can combine a communication device with a networking device that can inject (e.g., VPN router providing OSI Layer 2 access to an attacker) or modify (e.g., transparent firewall or similar device on the line) packets in the air-gapped network. These deployments can function without deployed malware, and therefore can be harder to detect but also can have less functionality. The most obvious scenario is establishing a foothold and communication channel inside the network of the systems, usually a switched Ethernet network.

Individual systems have their own set of controls and sensors that are directly connected to the system via a shared communication bus or point-to-point connections, which is commonly aggregated by a controller. Most of the controls or sensors can be replaced by compromised ones or a network intermediary device on the line that injects or modifies the data stream. The best point for the injection of a physical malicious control channel (or pre-programmed device or software) is the control board which allows the data stream of multiple inputs or outputs to be modified at the same time.

Without purposeful inspection, malware injected physical devices or breaches of air-gapping are hard to find as usually nobody is looking for those as long as everything functions properly. Therefore, the attacker can be dormant for months or years waiting for a specific moment or an opportunity to achieve an objective – disable the ship to get it stranded or take over control to crash it into another ship, the shore or a port.

The controller component (processor unit), such as ECDIS or MFD or a sensor, uses the multi-purpose cyber environment to test against cyber attacks. All the described potential attacks can be classified as either network (packet injection, modification, replay) or system (malware). This allows us to simulate and research all the possible outsider and even more diverse insider attacks.

4.2.3 Maritime-related experiences

The widely used legacy NMEA-0183 industrial protocol does not have sophisticated error correction, nor encryption; therefore, TCP packets containing NMEA-0183 messages can be faked or injected easily. A possible attack scenario could be a malware attack, where the malware is installed and sends TCP packets with valid sensor data. This kind of infection can be carried out in the supply chain or during software or chart updates on the ship.

Despite the fact that the periodicity of the sensor data is predictable, it is not checked in the ship's controller devices, so the displayed value can be overwritten if new -incorrect- data arrives right after the correct value. There are no resources describing how to detect abnormal behaviour in the messages with the valid payload. A possible solution could be an extension of an intrusion detection system that analyses and checks the data itself. This solution could also compare different sensor's data like speed data from speed sensor and location data from the GPS.

5. Summary and future works

The multi-purpose cyber environment introduced here aims to provide both education and research in cybersecurity issues for a specific vessel. Although this paper introduced only the technical and research perspective of the environment, education-related usage is the subject of a follow-up paper.

The solution increases cybersecurity research capabilities to include maritime cybersecurity, particularly for analysing systems, hardware and software components, protection development, and testing existing procedures.

Further development is needed to separate the ship-related and the simulator-related network traffic in order to create a more realistic environment.

References

- Bhatti, J. and Humphreys, T. E. (2017) Hostile Control of Ships via False GPS Signals: Demonstration and Detection, *NAVIGATION*, 64(1), pp. 51–66.
- Brinkmann, M. and Hahn, A. (2017) Testbed architecture for maritime cyber physical systems, in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*. *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Emden: IEEE, pp. 923–928.
- Caprolu, M. et al. (2020) Vessels Cybersecurity: Issues, Challenges, and the Road Ahead, *IEEE Communications Magazine*, 58(6), pp. 90–96.
- Davis, C. M. et al. (2006) SCADA Cyber Security Testbed Development, in *2006 38th North American Power Symposium*. *2006 38th North American Power Symposium*, Carbondale, IL, USA: IEEE, pp. 483–488.
- Drias, Z., Serhrouchni, A. and Vogel, O. (2015) Analysis of cyber security for industrial control systems, in *2015 International Conference on Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC)*. *2015 International Conference on Cyber Security of Smart cities, Industrial Control System and Communications (SSIC)*, Shanghai, China: IEEE, pp. 1–8.
- Explaining shipping* (2021). URL <https://www.ics-shipping.org/explaining/> (accessed 6.25.2021).
- Fortress Information Security (2020) *White Paper: Building A Sustainable Maritime OT Cyber Security Program*. URL <https://fortressinfosec.com/building-a-sustainable-maritime-ot-cyber-security-program/> (accessed 5.22.2021).
- Frank, M., Leitner, M. and Pahi, T. (2017) Design Considerations for Cyber Security Testbeds: A Case Study on a Cyber Security Testbed for Education, in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Orlando, FL: IEEE, pp. 38–46.
- Goud, N. (2018) *Cyber Attack on COSCO, Cybersecurity Insiders*. URL <https://www.cybersecurity-insiders.com/cyber-attack-on-cosco/> (accessed 5.17.2021).
- Gourlay, L. (2000) *Chambers guide to English for I.T. and the internet*. Edinburgh: Chambers.
- Heering, D., Maennel, O. and Vanables, A. (2020) Shortcomings in cybersecurity education for seafarers. Preprint. International Maritime Organization (2021) *Autonomous ships: regulatory scoping exercise completed*. URL <https://www.imo.org/en/MediaCentre/PressBriefings/pages/MASSRSE2021.aspx> (accessed 6.25.2021).
- Lund, M. S. et al. (2018) Integrity of Integrated Navigation Systems, in *2018 IEEE Conference on Communications and Network Security (CNS)*. *2018 IEEE Conference on Communications and Network Security (CNS)*, Beijing, China: IEEE, pp. 1–5.

- Pen Test Partners (2020) *Speed 2 – The Poseidon Adventure – Part One*. URL <https://www.pentestpartners.com/security-blog/speed-2-the-poseidon-adventure-when-cruise-ships-attack-part-1/> (accessed 7.19.2021).
- SAFETY4SEA (2019) Norsk Hydro lost about \$35-40 million after cyber attack, *SAFETY4SEA*, 1 April. URL <https://safety4sea.com/norsk-hydro-lost-about-35-40-million-after-cyber-attack/> (accessed 5.17.2021).
- Salunkhe, O. *et al.* (2018) Cyber-Physical Production Testbed: Literature Review and Concept Development, *Procedia Manufacturing*, 25, pp. 2–9.
- Smogeli, Ø. (2017) Digital Twins at Work in Maritime and Energy.
- Smogeli, Ø. (2017) FEATURE FEBRUARY 2017, p. 7.
- Tam, K., Forshaw, K. and Jones, K. (2019) Cyber-SHIP: Developing Next Generation Maritime Cyber Research Capabilities, in *Conference Proceedings of ICMET Oman. International Conference on Marine Engineering and Technology Oman*, Muscat, Oman: IMarEST.
- Tam, K. and Jones, K. (2018) Cyber-Risk Assessment for Autonomous Ships, in *2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, Glasgow: IEEE, pp. 1–8.
- United Nations (2021) *Review of Maritime Transport 2020*. S.l.: United Nations.
- WIRED (2018) *The Untold Story of NotPetya, the Most Devastating Cyberattack in History*. URL <https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/> (accessed 6.25.2021).

Ethical and Legal Aspects Pertaining to law Enforcement use of Drones

MM Watney

University of Johannesburg, South Africa

mwatney@uj.ac.za

Abstract: Law enforcement is an information-based activity. The use of drones (also referred to as unmanned aerial vehicles or UAVs) for policing may be beneficial as an aerial surveillance tool in gathering information pertaining to crime prevention, detection and /or investigation which are conducted in the interest of national security. In most instances, the law enforcement use of drones for purposes of search and rescue, crime scene investigation and hostage situations are not controversial. However, police use of drones for crowd monitoring and protests may be contentious as it may violate various human rights such as the right to privacy which includes data protection, free speech, right to protest and freedom of movement. These rights must be balanced against public safety. The discussion focuses on identifying ethical and legal concerns relating to the use of drones by the police and how these concerns should be addressed. It highlights that the danger is not the drone technology itself, but how it is used and the manner in which the police deal with, process and act upon information gathered, in order to prevent or control crime. The use of drone technology for surveillance impacts on human rights. There is a risk that surveillance may manifest itself in governmental domination and power if no safeguards are in place to curtail pervasive surveillance. It should be established whether domestic drone policing is in general so intrusive that the drawbacks outweigh the benefits of using it for public safety purposes. The manner in which these issues are addressed may serve as a guideline to countries who are considering the use of drones for law enforcement.

Keywords: police drones; law enforcement use of drones; legal and ethical aspects pertaining to domestic aerial policing; aerial drone police surveillance; constitutional impact of the use of police drones.

1. Introduction

Drones — also called Unmanned Aerial Vehicles (UAVs) - are simply described as aircraft that do not carry a human pilot and are remotely controlled (NCIS, 2021). It could be compared to a mobile CCTV camera.

Drones have been around for more than two decades. The use of drone technology started in the military. It's origin dates back to World War 1 when both the United States (US) and France worked on developing automatic, unmanned airplanes (Insider Intelligence, 2021). Since the mid-1990s, countries such as the US have been operating drones and utilizing armed and unarmed drones in combat situations (Dorsey and Amaral, 2021).

Although drone technology may have started in the military, the last few years have been significant in terms of drone adoption, usage expansion across industries and global awareness. Governmental and non-governmental actors are now moving towards the utilisation of drones. This discussion focuses specifically on domestic law enforcement (governmental actor) usage of drones. It is not directed at the use of drones for border control or for commercial purposes.

At the outset of this discussion it is important to take note that drones pose a security and safety threat. A drone is essentially a flying computer and its security systems are capable of being "hacked" or breached. Drones may also be used for malicious purposes such as to commit a crime. It could be used for physical attacks such as face-recognition drones or drones smuggling contraband (UNICRI, 2019). In 2021 it was reported that a prisoner in Italy used a drone to smuggle a gun into prison which he used to shoot fellow inmates (AFP, 2021). There is a possibility that a drone could be used for a terrorist attack (Police Executive Research Forum, 2020).

It is suggested that the advantages of drone usage for policing outweigh the security and safety threats. Drones are ideally suited for numerous aerial law enforcement tasks such as search and rescue, crime scene investigation, hostage situations or crowd monitoring to detect any suspicious-looking members during an event. Drone technology provides a better field of view and it may enter areas that would be perilous to people (Police Executive Research Forum, 2020). It may be equipped with facial recognition technology or with license plate readers. Law enforcement agencies argue that aerial surveillance, via a drone or other means, is less intrusive and less likely to result in deadly altercations with a distrustful public than controlling on foot. Deploying a police drone protects a police officer from physical danger (Enemark, 2021). It is also much less expensive and complex to operate than a manned helicopter (Greenwood, 2020).

The downside to police use of drones is public distrust. Some members of the public may equate it to military drones and perceive it as the militarisation of the police, whereas others may perceive the use of police drones as a serious violation of the right to privacy and data protection.

Drone technology may be seen as a tool - and one of many - for improving policing. Visible policing by means of drones may address the public's demand for safety and security. Drones undeniably have the potential to be an asset to law enforcement to ensure public safety, but there are ethical and legal aspects that must be considered.

The looming prospect of expanded use of drones (not only by law enforcement but also by non-governmental actors), has raised understandable concerns for lawmakers in countries in which it is used. Countries that are not yet making use of drones may learn from the challenges countries experience where law enforcement are using drones.

2. Identifying ethical and legal aspects pertaining to law enforcement use of drones

The 2020 George Floyd protests against police brutality in the US did not only constitute the biggest wave of protests in the US since the 1960s, but they were also the most monitored protests in American history. The law enforcement surveillance of the Black Life Matters (BLM) protest brought about many legal and ethical issues taking into consideration that citizens have a right to peaceful protests and free speech. The question asked pertaining to the protests is how the police justifies the use of the many surveillance tools - from social media analytics to body cameras and drones - to monitor the protests and protesters (Greenwood, 2020). It could be argued that surveillance was used for national security purposes to ensure public order and/or prevent and/or detect crime. National security may be described as the capacity of a state to provide for the protection and defence of its citizenry. On the other hand, how is a balance achieved between the protection of human rights, such as free speech, privacy which includes data protection and national security against excessive governmental power and domination?

The discussion specifically focuses on the use of drones for policing. It focuses on various interlinked issues relating to the ethical and legal use of drones for law enforcement purposes, such as

- whether specific legislation should regulate the use of drones for policing (in addition to aviation legislation);
- whether the use of drones amount to search;
- whether a warrant must be required before the police may use drones;
- whether drones may be weaponised;
- the admissibility of drone gathered evidence (such as videos) in a court;
- the manner in which it impacts on privacy and data collection; and
- the risk of pervasive surveillance.

3. Overview of the legal position in countries pertaining to police use of drones

3.1 Introduction

An overview of countries in respect of drone policing serves as a background to the discussion of the ethical and legal concerns relating to domestic drone policing.

3.2 The U.S.

The U.S. police have been using drones since approximately 2016 (Police Executive Research Forum, 2020).

The Federal Aviation Administration (FAA) is the designated authority to regulate aircraft, including drones, that fly in U.S. airspace (Police Executive Research Forum, 2020). The FAA also currently prohibits anyone, including law enforcement, from attaching firearms or similar weapons to drones. In 2015 North Dakota became the first state to legislate in favour of drones armed with non-lethal weapons whereas several other states have expressly prohibited any kind of drone weaponization (Enemark, 2021; see par. 4.6 hereafter).

The mission of the FAA is "to provide the safest, most efficient aerospace system in the world" (FAA, 2021). This mission does not include regulating privacy-related issues and the FAA has no guidelines regarding privacy (Fischer et al, 2017).

The Fourth Amendment of the U.S. Constitution is considered the bedrock of privacy as it protects the right of people to be secure against unreasonable searches and seizures. The ultimate goal of this provision is to protect people's right to privacy and freedom from unreasonable intrusions by the government. The Fourth Amendment right is therefore relevant in respect of the use of drones for aerial surveillance.

Several Supreme Court judgements deal with aerial surveillance, although not specifically by means of drones. The Supreme Court's aerial surveillance jurisprudence established principles pertaining to privacy and the Fourth Amendment that are applicable to law enforcement use of drones.

In *Katz v United States* (1967) the right to privacy was extended to protect people in places in which people have a reasonable expectation of privacy. It is unconstitutional under the Fourth Amendment to conduct a search and seizure without a warrant anywhere that a person has a reasonable expectation of privacy, unless certain exceptions apply (Fischer et al, 2017; Feeney, 2021). The Supreme Court found in *United States v Knotts* (1983) that individuals do not have an expectation of privacy when driving a car on public roads (Fisher et al, 2017). With reference to the *Knotts* case, it may be found that drones may surveil moving targets and continue to track their location in public.

Thus, a person has a legitimate expectation of privacy if he honestly and genuinely believes the location under surveillance is private, and if a reasonable person in the same or similar circumstances would believe the location to be private as well. Therefore, the government has more latitude to legally survey communications in a public place than it does in a private place. This may explain the use of drones in respect of the 2020 BLM protests.

Aerial observations of the outside of a home are generally not prohibited by the Fourth Amendment, but the government must conduct the surveillance from a public navigable airspace in a non-physically intrusive manner. The government conduct may not reveal intimate activities traditionally associated with the use of the home (McNeal, 2014). In *California v Ciraolo* (1986), the Supreme Court ruled that the defendant's yard was not constitutionally protected from observation from a public vantage point such as public airspace and the warrantless aerial surveillance did not violate the Fourth Amendment. The ruling in *Ciraolo* could be interpreted to allow unrestricted usage of drones in public airspace over any surface (Fischer et al, 2017).

In *Florida v Riley* (1989) the Supreme Court held that aerial surveillance above 400 feet does not constitute a search and does not require a warrant (Fischer et al, 2014). McNeal (2014) indicates that by tying the Fourth Amendment's protections to the location in airspace from which the surveillance was conducted, the Supreme Court has left open the possibility that low altitude surveillance may violate the Fourth Amendment. The drones that law enforcement are most likely to acquire and operate are most effective at altitudes below 500 feet. McNeal (2014) recommends that legislators should follow a property rights approach to aerial surveillance, explicitly extending to property owner's rights in their airspace up to 350 feet above ground level. Such an approach may solve most public and private harms associated with drones by allowing the landowner to exclude intrusions into their airspace by government and private parties.

In *Kyllo v United States* (2001), the Court addressed the constitutionality of using technology to survey the inside of the defendant's home without actually entering the home and found that this search constituted a Fourth Amendment search and would be presumptively unreasonable without a warrant. As a result of *Kyllo*, the government cannot equip drones with devices that can see inside a home without a warrant (Fischer et al, 2017).

Feeney (2021) makes a valid point by emphasizing that although Supreme Court precedent may not require warrants for drone surveillance, state law makers may improve upon this standard. He is correct when he states that "the Supreme Court precedent sets the floor, not the ceiling." (Feeney, 2014). This explains why most states provide that the usage of a drone by law enforcement must be executed by means of a warrant subject to exceptions, namely extreme circumstances, such as an imminent danger to life or property (Fischer et al, 2017; NCSL, 2021).

3.3 United Kingdom (UK)

Police in the UK embrace the use of drones. In November 2020 it was revealed under the Freedom of Information Act (PFA) 2012 that 40 of the 43 national police forces had adopted drone technology, with over 288 drones operational across the board (Jackson, 2021).

Drone operations are conducted in compliance with the relevant legislation and are overseen by the Civil Aviation Authority (CAA). The police do not need special permission to fly in UK airspace as the police work is classed as “commercial gain” and therefore permission is not required (Jackson, 2021). Since the use of drones by the police is considered as a commercial operation, police are allowed to fly a drone in congested areas and permitted the overflight of people and/or premises, provided other conditions and minimum separation distances are met. Police officers must abide by the same regulations as other drone pilots in the UK. The drones must follow all CAA regulations and fly within 500 m of the pilot. The drones cannot fly outside of the pilot’s line of sight unless a second pilot is used. Drones are restricted from flying above 400 ft (Jackson, 2021).

As drones are considered the equivalent of a mobile CCTV, police must comply with the Surveillance Camera Code of Practice (available at <https://www.gov.uk/government/publications/surveillance-camera-code-of-practice>) which was issued under section 30 of the Protection of Freedoms Act 2012. It provides guidance on the appropriate use of surveillance camera systems in public places by the police in England and Wales. The police has a legal duty to adhere to the principles of the Code when operating overt surveillance camera systems in public places. When it comes to data collection, processing and retention, the police must comply with the Data Protection Act 2018 (as updated on 1 January 2021).

3.4 France

France is an European Union (EU) member state. Strict privacy and data protection laws are applicable in EU member states. The use of police drones for surveillance of demonstrations was banned in 2020 (see par. 4.4 hereafter). In 2021, parliament was presented a bill on criminal responsibility and internal security, intended to provide a long-awaited legal framework for the use of drones by law enforcement agencies (Pollet, 2021). The bill defines the scenarios in which police may use drones. Those are limited to preventing attacks on people and property, fighting terrorism, traffic regulation, border surveillance and assemblies likely to result in serious disorder (Pollet, 2021).

3.5 South Africa (SA)

The SA police service is not yet using drones for law enforcement purposes, but it will soon be utilised. In 2020 representatives from the South African Police Service (SAPS) and Johannesburg Metro Police Department (JMPD) stated that drones will be used in the future for law enforcement, public safety and disaster management (Cronje, 2020). In 2021 a private security group indicated that they would become the first security providers to use drones to track criminals in specific residential areas (Staff writer, 2021).

4. Discussion of the legal and ethical aspects pertaining to police use of drones

4.1 Introduction

The discussion hereafter explores ethical and legal aspects with respect to police use of drones. Each of the aspects could warrant a discussion on its own. The purpose of this discussion however, is to provide an overview of the various aspects and to establish whether the concerns relating to the legal and ethical aspects are of such a nature that it outweighs the benefits derived from the use of police drones.

4.2 Purpose-specific regulation

Drones are regulated by aviation legislation and regulations which address issues such as the design, manufacturing of drones and whether a drone must be registered etc. The question is whether there should be a purpose-specific act that, in addition to the aviation legislation and regulations, regulate the use of drones by police. Such an act will address various legal and ethical issues specific to police use of drones, such as the circumstances in which the police may use a drone, warrant requirement, privacy and data protection, the use of weaponised drones and the admissibility of drone evidence in court proceedings.

In the US, many states have purpose-specific legislation. The US government could on federal level consider a uniform policy in respect of police drone usage. In 2021, France introduced legislation governing police drone use. Such purpose specific legislation may contribute to legal certainty, transparency and accountability.

4.3 Warrant requirement

Some privacy advocates have called for an outright ban on drone usage, but such a call is not supported as this will not be to the benefit of the public safety in general as there are many instances in which drones may assist law enforcement.

Many privacy advocates have focused solely on requiring warrants before the use of drones by law enforcement to counter-act the threat of surveillance. The use of drones by means of a warrant is supported, but there may be instances where it should be allowed without a warrant. In this regard, McNeal (2014) makes a valid argument that legislators should reject calls for a blanket requirement that all drone use must be accompanied by a warrant. There may be circumstances where the police urgently need to use a drone for aerial surveillance and there may not be time to apply for a warrant, for example in a a hostage situation or a protest turned violent. McNeal's (2014) criticism of a blanket requirement of a warrant in all cases is justifiable as such a technology-centric approach harms public safety and deprives law enforcement of a tool that they could use to protect people in public spaces.

4.4 Protests and crowd monitoring

Using police drones and other surveillance tools during protests and crowd monitoring may be controversial as most countries' citizens have a constitutional right to free speech and peaceful protesting (see par. 4.4 hereafter).

In 2020 the French Supreme Administrative Court ruled that drone surveillance of demonstrations may not be used (Monrey, 2020). Data protection law in France

- requires proof of the necessity and proportionality of a surveillance device, without which its use cannot be authorised; and
- provides that political opinions, as expressed at protests, are a form of "sensitive data" and a surveillance device must therefore be "absolutely necessary" in order for it to be legitimately deployed.

The Court found that there was no evidence that the drone surveillance was absolutely necessary to maintain public safety in crowds (Monrey, 2020). Banning the police from using drones does not eliminate the risk of, for example, a terrorist using a drone for an attack and the police should be equipped to deal with such a possibility.

Criticism was levelled against the use of a Predator-class drone during the U.S. 2020 BLM protest as it could be considered as "unduly intimidating and could have an unwelcome chilling effect on participation in public life" (Enemark, 2021). The protests – like most protests - were highly public events which were televised, it took place on streets where there were surveillance cameras and spectators photographing the event. The use of the drone was but one of many surveillance tools used and it provided information of possible threats to public safety.

In the UK, the police have used drones to monitor political protests, including those held by the non-violent BLM movement (Dodd, 2021).

South Africa has experienced many demonstrations that have turned violent. In early July 2021, South Africa experienced its worst riots and looting since 1984 (IOL reporter, 2021). In my opinion, the use of drones would have contributed to more visible policing during these riots and looting and would have improved law enforcement by ensuring the safety of the public.

4.5 Admissibility of drone-gathered evidence

Each drone has a memory card that captures images. Once the drone is back on the ground, the card is removed and the images can be transferred to a police official's computer. The vast majority of drone-gathered evidence will be in the form of photographs, video and audio recordings. This evidence must be gathered in accordance with a warrant (discussed at par. 4.3) and data protection laws must govern the processing, storage and access of the information (see par. 4.7).

4.6 Weaponised drones

It should be established whether the police may use a weaponised drone against a suspect to protect public safety, for crowd control or to disperse a riot.

Enemark (2021) discusses the ethical use of armed drones to enforce domestic law enforcement. Although the military uses armed drones, it cannot be compared to the domestic use of drones. At present the police have not deployed armed drones. However, it should be determined whether circumstances exist that could justify the police use of an armed drone. It could be argued that if a gun man goes on shooting spree, the killing of the perpetrator by means of an armed drone could be considered to prevent the loss of further lives. The police will only use force to mitigate very serious threats such as the potential loss of life. By using an armed drone in such

circumstances, the criminal threat would be neutralised with little exposure of danger to the police and members of the public.

In 2016, the UN Human Rights Council published a report on the proper management of assemblies (Enemark, 2021). It recommended that less lethal weapons be used to allow for the de-escalation of tensions. It also urged great caution in the use of weapons that are remotely controlled as would be the case with an armed police drone.

The issue of using an armed drone in the policing of a protest that has turned violent may be problematic. Drones armed with non-lethal weapons in the form of rubber bullets or water cannons could be deployed to de-escalate the violence (Enemark, 2021). Enemark (2021) warns however that there is a risk that the use of police drones forcefully against people rioting could have the opposite effect by aggravating an already angry crowd and escalating the overall threat to public safety. In addition, police drones could also cause panic and result in a stampede.

One can only speculate on the outcome of the use of weaponised drones pertaining to violent protests. The July 2021 rioting and looting in South Africa may have been contained by drones armed with non-lethal weapons, but as Enemark (2021) surmises, it could have intensified the situation.

4.7 Constitutional impact of domestic drone policing

The right to privacy is afforded stronger protection within the home than in public. For example, a protester in a public space will still have a right to privacy albeit a much weaker right than in his home. This right must be weighed against public safety.

The right to privacy includes data protection. As indicated, the police collect large amounts of information for the prevention, detection and investigation of crime. There must be safeguards in place to protect the collection, processing and sharing of data. The United Nations Inter-regional Crime and Justice Research Institute (UNICRI) propose that law enforcement agencies comply with the following requirements regarding data protection, namely fairness, accountability, transparency and explainability (UNICRI, 2019).

The U.S. does not have federal law that provides for data protection and this constitutes a major difference between the approach of the US and EU to data protection. As indicated, the EU provides for strict data protection. Law enforcement agencies must show that their systems, technology and processes are fully compliant with the General Data Protection Regulation (GDPR). The Law Enforcement Directive (referred to as LED) is a piece of EU legislation, parallel to the GDPR, which deals with the processing of personal data by data controllers for law enforcement purposes which fall outside of the scope of the GDPR (CNIL, 2021). The GDPR and the LED are both part of the European data protection package (CNIL, 2021). In 2020, South Africa implemented the Protection of Personal Information Act 4 of 2013 which is based on the GDPR. As indicated, the UK police must comply with the Data Protection Act. In respect of protests, the UK police may interfere with the right to privacy if such interference is proportionate to the objectives of maintaining public order and preventing or detecting a crime (UK Privacy International, 2021). The police may only process data regarding a protest that are necessary for the administration of justice. When it comes to facial images, the processing must be strictly necessary. Legislation provides that the police may hold data only as long as it is necessary (see UK Privacy International, 2021).

4.8 Drones as a surveillance tool

Drone technology is one of many types of surveillance technologies that police use (Greenwood, 2020). The use of police drones for surveillance is an example of visible policing and not a covert surveillance tool. As indicated, drones should not be used at random as a general surveillance tool, but used for a specific purpose to promote public safety and in circumstances that warrant the use of a police drone. The police use of drones must be conducted in accordance with a warrant subject to exceptions and the data collected must comply with data protection laws.

The 2013 Snowden disclosures and 2018 Facebook-Cambridge Analytica scandal illustrated that the risk of pervasive surveillance is real (Watney, 2021). The danger exists that surveillance technology may be abused and result in an Orwellian big brother society and ultimately a police state, especially in circumstances where mass

surveillance is conducted. It is therefore important that police use of drones comply with human rights safeguards (see par. 4.7).

5. Conclusion

The discussion illustrates that the use of drones by the police is beneficial for crime detection, prevention and investigation in specific circumstances. The legal position of countries pertaining to police drone use provides a background to the legal and ethical aspects referred to. The concerns with the legal and ethical aspects must be addressed by means of safeguards to ensure that the police do not abuse their power. Fischer et al (2017) conclude that imposing safeguards in respect of domestic drone policing “would allow us to incorporate this new technology and separate the benefits from the dangers it presents.” By having clear guidelines in place for accountability and transparency, public distrust in the use of police drone technology will be assuaged. It is important that the police are equipped with the latest technologies, such as drone technology, in tackling crime and ensuring public safety, especially as criminals may use these technologies for crime commission.

References

- AFP. (2021) “Inmate thought to have used drone to smuggle gun into Italian prison”, [online], <https://www.euronews.com/2021/09/20/inmate-thought-to-have-used-drone-to-smuggle-gun-into-italian-prison>.
- California v. Ciraolo*, 476 U.S. 207 (1986).
- CNIL. (2021) “Law Enforcement Directive: what are we talking about?” [online], <https://www.cnil.fr/en/law-enforcement-directive-what-are-we-talking-about>.
- Cronje, J. (2020) “SAPS and JMPD see drones in their future”, [online], <https://www.defenceweb.co.za/aerospace/unmanned-aerial-vehicles/saps-and-jmpd-see-drones-in-their-future/>.
- Dodd, V. (2021) “Drones used by police to monitor political protests in England”, [online], <https://www.theguardian.com/uk-news/2021/feb/14/drones-police-england-monitor-political-protests-blm-extinction-rebellion>.
- Dorsey, J. and Amaral, N. (2021) “Military drones in Europe : Ensuring transparency and accountability”, [online], <https://www.chathamhouse.org/sites/default/files/2021-04/2021-04-30-military-drones-europe-dorsey-amaral.pdf>.
- Enemark, C. (2021) “Armed drones and ethical policing: risk, perception and the tele-present officer”, *Criminal Justice Ethics*, 2021 vol 40, no 2, pp 124 – 144, [online], <https://www.tandfonline.com/doi/full/10.1080/0731129X.2021.1943844>.
- Feeney, M. (2021) “Does the 4th Amendment prohibit warrantless drone surveillance”, [online], <https://www.cato.org/blog/does-4th-amendment-prohibit-warrantless-drone-surveillance>.
- Federal Aviation Administration. (2021), [online], <https://www.faa.gov/about/mission/>.
- Fischer, A.S, Cartmell, J.E. and Frank, L. (2017) “Drones, a new front in the fight between government interests and privacy concerns”, [online], <https://www.iadclaw.org/defensecounseljournal/drones-a-new-front-in-the-fight-between-government-interests-and-privacy-concerns/>.
- Florida v. Riley*, 488 U.S. 445 (1989).
- Greenwood, F. (2020) “How to regulate police use of drones”, [online], <https://www.brookings.edu/techstream/how-to-regulate-police-use-of-drones/>.
- Insider Intelligence. (2021) “Drone technology uses and applications for commercial, industrial and military drones in 2021 and the future,” [online], <https://www.businessinsider.com/drone-technology-uses-applications?IR=T>.
- IOL reporter. (2021) “Live updates#SayNoToLooting – violence, looting continues despite SANDF deployment”, [online], <https://www.iol.co.za/news/south-africa/gauteng/live-updates-saynotolooting-violence-looting-continues-despite-sandf-deployment-1f4e0b9c-8c88-494e-8160-aeacff164ec1>.
- Jackson, B. (2021) “ Drone solutions for the police: How are the police using drones?” [online], <https://coptrz.com/drone-solutions-for-police-how-are-the-police-using-drones/>.
- Katz v. United States*, 389 U.S. 347 (1967).
- Kyllo v. United States*, 533 U.S. 27, 29 (2001).
- McNeal, G (2014) “Drones and aerial surveillance: considerations for legislatures”, [online], <https://www.brookings.edu/research/drones-and-aerial-surveillance-considerations-for-legislatures/>
- Monrey, M. (2020) “Court in France bans police drones at demonstrations”, [online], <https://digit.site36.net/2020/12/22/court-in-france-bans-police-drones-at-demonstrations/>.
- NCSI (National Conference of state legislatures). (2021) “Unmanned Aircraft State Law Landscape”, [online], <https://www.ncsl.org/research/transportation/current-unmanned-aircraft-state-law-landscape.aspx>.
- Police Executive Research Forum. (2020) “Drones: A Report on the Use of Drones by Public Safety Agencies—and a Wake-Up Call about the Threat of Malicious Drone Attacks”, [online] <https://www.policeforum.org/assets/Drones.pdf>.
- Pollet, M. (2021) “France’s parliament consider the use of drones by police, again.” (2021) [online], <https://www.euractiv.com/section/data-protection/news/frances-parliament-considers-use-of-drones-by-police-again/>.

Staff writer. (2021) "Security company is now using drones to track criminals in south Africa's suburbs", [online], <https://businesstech.co.za/news/technology/493193/security-company-is-now-using-drones-to-track-criminals-in-south-africas-suburbs/>.

Terry v. Ohio, 392 U.S. 1 (1968).

UK Privacy International. (2021) "Fact sheet on your data rights in relation to police surveillance at protests", [online], <https://privacyinternational.org/long-read/4507/fact-sheet-your-data-rights-relation-police-surveillance-protests>.

United Nations Interregional Crime Research Institute (UNICRI). (2019). "Artificial Intelligence and robotics for law enforcement", [online], <https://www.europarl.europa.eu/cmsdata/196207/UNICRI%20-%20Artificial%20intelligence%20and%20robotics%20for%20law%20enforcement.pdf>.

United States v. Knotts, 460 U.S. 276 (1983).

Watney, MM. (2020) "Mobile Phone Surveillance: An Overview of Privacy and Security Legal Risks", Proceeding of the 20th European Conference on Cyber Warfare and Security (ECCWS 2021) A Virtual Conference hosted by University of Chester, UK on the 24th-25th June 2021, pp 462 – 469.

Social Media Privacy Using EDEE Security Model

Benjamin Yankson, Eric Cajigal Delgado, Amjad Al-Jabri, Natalie Gitin and Sydney Davidson
University at Albany, USA

Byankson@albany.edu

Abstract: Social Media platforms have become a significant part of our daily lives and a modern way to connect friends and family, document our lives, and share other great personal information about our lives. These activities leave us vulnerable to privacy and security breach due to lapse security controls necessary to protect users' sensitive data on these platforms. We conducted exploratory privacy and security analysis on paramount social media platforms such as Facebook and Snapchat and determined that current Social Media privacy and security posture insufficient and proposed Social Media platform Security through "Educate," "Determine," "Enable," and "Evaluate" (EDEE) Security model to address the evolving Social Media platform security as a growing concern in Cybersecurity for individual using the platform and companies hosting the platform.

Keywords: Security, Social Media, Privacy, Facebook, Snapchat, Model

1. Introduction

Social Media platforms allow us to interact with people around the globe with a click of a button and have become a crucial part of everybody's life. These platforms allow us to call, text, share videos and pictures, share experiences, and do much more at any time or any place for free. However, there is a price of once privacy at stake considering the information, we share on Social Media platforms without any assurances that platforms are putting in effort in protecting our intimate details against hackers or misuse (Center, 2020). Some of these large corporations, who own these platforms, have taken advantage of users' insatiable appetite to share intimate information about their lives on their platforms. Yet, they have made no serious attempt to secure user information. Also, there have been instances where these companies have been complicit in misusing users' data. For example, On March 16, 2018, Facebook admitted illegally transferring 50 million user profiles to Cambridge Analytica, the data analytics firm that harvested the data without user consent (Center, 2020). Most everyday users have always assumed that buying and selling data on the Internet was only done by hackers and cyber-criminal and not the legitimate organizations who runs these platform (Yankson et al., 2019). As a result, some platforms like Facebook and Snapchat have faced the United States courts many times concerning inadequate data protection or data exploitation resulting in a user privacy breach. Several of these lawsuits have resulted in settlement scenarios where these platforms have been forced to adopt new terms and conditions to improve their security posture. However, in many instances, there have been no meaningful actions to protect user data or privacy (Viala, 2018) (CBS, 2010).

As per Gemalto (Viala, 2018), a global leader in digital security, in their 2018 worldwide database breach index findings, the first half of 2018 experienced approximately 4.5 billion data records breach in 945 separate attacks. Out of these attacks, only 6 were Social Media platform attacks; yet, they accounted for more than 56% of the overall data record exposed (Viala, 2018). Including in the Gemalto breach index is the infamous Cambridge Analytica-Facebook privacy breach. About 87 million Facebook user data was shared without users given the opportunity to give consent or deny consent. Table 1 presents some very well-known Social Media platform data breaches with the affected Social Media platform, the year of the breach, and the number of impacted users account as evidence of the criticality of addressing privacy and security concerns in social media user. Although while Social Media platform privacy and security breaches impacts may not seem serious to some people, based on the presumed type of information posted; the data available in the vast majority of Social Media user accounts can be leveraged by a malicious attacker to conduct social engineering attack possibly leading to identity theft. Further, access to such user information can be leveraged to access even users' other personal online or employer-related authentication credentials, resulting in data breaches such as users' email, employer payroll information, or customer databases.

Table 1: Well Known Social Media Platform Data Breaches

Social Media Platform	Breach Date	Impact (users)
Facebook	Sept 2019	419,000,000
500px	Feb 2019	14,800,000
Myspace	2016	360,000,000
Google+	Dec 2018	52,500,000
Instagram	Sept 2017	6,000,000
Facebook	Sept 2018	50,000,000

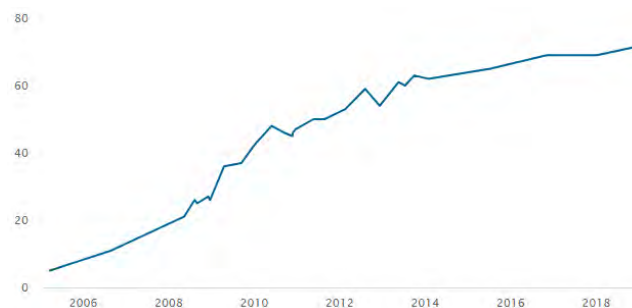
Based on the current security and privacy of Social Media platform breaches, as depicted in Table 1, some users can finally become aware that Social Media platforms or associated applications abuse their data. We hypothesize that the current security controls on most of these platforms are inadequate, and the gaping privacy and security holes on these platforms or related applications can lead to serious privacy and security risk. As a result, we present a theoretical research analysis of current Social Media privacy and security posture by evaluating some key platforms. Based on the results, we present the Social Media platform Security through Educate, Determine, Enable, and Evaluate (EDEE) Security model. EDEE improves the evolving Social Media platform security as a growing concern in Cybersecurity for individuals using these platforms and companies hosting them. This research work is divided into Section 2: analysis of the background and related work. Section 3: focuses on methodology and Facebooks' and Snapchat privacy and security analysis case study. Section 4: focuses on the EDEE security model. Section 5: Summarizes the work in the Conclusions.

2. Background and Related Work

2.1 Social Media – Past to Present

Social Media platforms have only been around for a little more than two decades (Lamdan, 2015). Since its inception, we have witnessed conceptions, shutdowns, and those that have turned into obscurity, leading to a quiet and forgettable demise. In 1999, the portal was bought from the original developer by YouthStream, Media Networks, and it was shut down (Jones, 2020). As time went on and more people gained access to computers conjunctively with the creation of the Internet, there was bound to be a place for all these people to connect and socialize through an electronic medium leading to the rise of social media platforms. The Internet is now home to a few dozen Social Media platforms built solely to connect people online and made to foster a sense of community irrespective of geographical location restriction or language. Today, we have Facebook, Twitter, Instagram, Snapchat, YouTube, Tumblr, TikTok, Pinterest, LinkedIn, Discord, and GroupMe. All these platforms have a unique interface, and it's intended to provide us with an opportunity for purposeful communication in our personal and professional lives. Millions of people use these platforms daily, and it is up to the companies that build and maintain these platforms to ensure that they are secure and protect user privacy. It is important that Social Media companies see security as more important than anything considering the significant amount of user data in their respective care.

According to the PEW research group (PEW, 2020), as of Feb 2019, 72% of United States adults have at least one Social Media site account. Figure 1 presents a graph of Social Media usage growth from 2006 to 2018, demonstrating radical usage growth. It exhibits a very noticeable rapid increase in Social Media usage over the 12 years. There was a steep incline in 2008, and it continued that trend until it began to slowly flat line in 2018.

**Figure 1:** Trends of U.S Adults with a Social Media Account (PEW, 2020)

The reason why the usage of Social Media increased so much between 2008 and 2018 was that apps like Instagram and Snapchat were just being released in 2010 and 2012, respectively (McIntyre, 2014). These two apps and many others (Vine, Pinterest, etc.) were released within this period, which spiked the use of Social Media apps. Additionally, we can also attribute this growth to smartphone technological advancement and increased ease of access to these devices over time, thus making it easier for individuals to use these platforms more easily. Before this boom, Social Media platforms, particularly Facebook, were accessed by computers rather than smartphone devices.

2.2 Development of Security

According to the United States Department of Defense (DoD), Social Networking Sites (SNSs), such as Facebook and Snapchat, remain vulnerable to web application attacks, web browser vulnerabilities as well as social engineering attacks (Waldyogel, 2017). Amongst the web application attacks available today, the DoD specifically notes a susceptibility to "Buffer Overflow Attacks," "Cross-Site Scripting Attacks," and "Code Injection Attacks" (Waldyogel, 2017). Buffer Overflow attacks happen when an attacker injects malicious code into a program that runs beyond the allotted amount of memory. In turn, this causes the system to crash, thereby making it vulnerable to compromise. Code Injection Attacks occur when an attacker utilizes an ability "to inject malicious code into a system that" can then be executed by an application (Waldyogel, 2017). Finally, a Cross-Site Scripting Attack is a "type of code injection that occurs in the form of a browser-side script" (Waldyogel, 2017). This type of attack can result in sensitive data exfiltration and the possible destruction of the affected system. Common amongst all three types of attacks is the ability to exploit vulnerabilities on these platforms or vulnerabilities found on the targeted system to establish a compromised asset. These attacks can be utilized against flaws found in Operating Systems and software packages. DoD mentions that web browser vulnerabilities exploited while using a social media platform specifically pertained to publishing content in the form of plain text, HTML, or active content such as JavaScript or Adobe Flash (Waldyogel, 2017).

The results of a malicious payload can come in the form of downloaded malware, an HTTP request to a malicious site, or even a Denial-of-Service (Waldyogel, 2017). The DoD attributes social engineering methods, specifically "Phishing" attacks, as one of the more elusive attack vectors. A Phishing attack is a type of attack in which a falsified email is sent to the target to gain information from them, such as login credentials. This can also result in the user downloading a malicious file that appears to be legitimate. These types of attacks typically "do not flow their network email servers" and, thereby, "can escape implemented email content filters" (Waldyogel, 2017). These platforms have been hacked on multiple occasions, and the portals have evolved as a response. For example, when you log in to Facebook after some time or from an unknown location, the website makes you go through a test to make sure you are who you say you are.

3. Methodology

For this study, we select to conduct an exploratory privacy and security analysis on paramount social media platforms such as Facebook and Snapchat; to determine if current Social Media privacy and security posture is insufficient based on select cases, existing literature, and perceived privacy. Contrary to the existing null hypothesis that users privacy is protected on Social media, we hypothesize that:

H₁: The current security controls on most of these platforms are inadequate. The gapping privacy and security holes on these platforms or related applications can lead to serious privacy and security risk.

3.1 Facebook Case Study Analysis - Perceived Privacy

In 2003, a Harvard sophomore hacked into the University's student database to create a social media webpage called "FaceMash" (Huggman, 2015), which was later shutdown. In 2004, the creator of FaceMash used his knowledge and experience to make a Social Media site for Harvard students called "Facebook." The use of Facebook spread out of the boundaries of Harvard to more mainstream use. As of December 2019, Facebook hit the highest record of users. During the third quarter of 2012, Facebook active users surpassed one billion, making it the first social network to pass that mark, with approximately roughly 2.91 billion monthly active users as of the fourth quarter of 2021 (Statista, 2021). Facebook has become the most popular Social Media platform of its time. Although Facebook user growth is significant, based on Table 2, we can fairly indicate that the organization can do more to improve on security and privacy practices in order to protect users' privacy. Table 2 presents a decade of major incidents resulting in a data breach that impacted Facebook user privacy.

Table 2: A decade of data breaches impacted Facebook user privacy (Heiligenstein, 2021)

Date	Incident	Privacy Impact
April 2021	Facebook users' data was leaked on an online forum	Over 530 million user data was posted in an online hacking forum.
June 2020	Facebook accidentally made available users' information to third-party developers	Users' personal information was made available without consent.
March 2020	Hacker group captures data from Facebook Accounts	42 million more users were impacted
December 2019	Hacker group captures data from Facebook Accounts	Approximately 267 million accounts account impacted
September 2019	Data for Facebook users found on an exposed server	419 million Facebook Users impacted
April 2019	Facebook uploads users' Email contacts without permission	1.5 million users impacted
April 2019	Facebook user records found on public Server	Approximately 540 million Facebook user records captured by app developers stored in an Amazon cloud public server
March 2019	Facebook passwords stored in plaintext files exposed to employees, some dating back to 2012.	600 million Facebook user passwords had been stored in plaintext files
December 2018	New York Times discovers Facebook is sharing user data without permission to sell users' information to over 150 companies.	All Users
September 2018	Attackers access data of up to 90 million Facebook Users	50 to 90 million users impacted
May 2018	Facebook bug makes 14 million Users' private posts Public	14 million users' private posts were shared publicly even though they were initially posted with viewing limitations
March 2018	Cambridge Analytica exploited a loophole in Facebook's API that allows compiling profile data not just from users who downloaded the app but also from their friend networks	50+ million users impacted
June 2013	Bug exposes personal data of 6 million Users	approximately 6 million users sensitive personal data impacted

As per the April 2021 incident depicted in Table 2, over 533 million Facebook users' personal information, including phone numbers, was made available online (Heiligenstein, 2021). Similarly, in December 2019, approximately 267 million Facebook user accounts data, including names, phone numbers, and Facebook ID, was found unprotected on the dark web. Following the 2019 discovery, in March 2020, a second discovery was made on a server containing about 42 million more users, bringing the total up to 309 million. (Heiligenstein, 2021). In July 2019, the FTC fined Facebook \$5 billion for privacy violations and mandated privacy requirements to align Facebook user privacy requirements. Based on these occurrences, as illustrated in Table 2, Facebook's user data privacy, security, and privacy rights while using the application show inadequacy.

Adding to privacy issues presented in Table 2, In June of 2020, Facebook accidentally shared user data with a third-party company known as "Cambridge Analytica." Before the Cambridge Analytica incident was unearthed, Facebook users were unaware that Cambridge Analytica was taking their information. A lot of the time, when another application asks for the user's permission to access Facebook, the user may just hit "accept" without reading any of the fine print. It can be very dangerous because sensitive information could be taken, such as bank account information, where the user lives, and this data could hurt them financially or physically. In addition, Facebook users, in most cases, are unaware of relinquishing their intellectual property of the content they post or private information shared. In addition to the Cambridge Analytica data breach, another incident targeted Facebook and affected over 90 million users housed on the social media platform giant (Center, 2020) (Jones, 2020). The attack vector that led to the result of the breach ended up allowing the hackers to obtain access tokens for Facebook users. The access tokens are the digital keys that keep users logged in to Facebook without asking them to re-enter their password every time they use the website. The attack was a combination of 3 bugs or software flaws, which created this vulnerability. These bugs were the "View As" feature, the code that lets people with the user a happy birthday, and the video uploader, which would create an access token even though it should not have been able to by design. To address this issue, Facebook reset all the access tokens without knowing the hackers' intentions. Nathaniel Gleicher, Facebook's head of Cybersecurity Policy, stated that it is unclear who was behind the attack and commented: "sophisticated adversaries" (Kerner, 2018).

One thing that stands out in that snippet from the Facebook Terms and Conditions is the line “consistent with your privacy and application settings (Sayin et al. 2019). This implies that there are ways to turn this off, so to say, but the user would have to know how to do this. As explained by John Quain, Minute Technology consultant, during a CBS news segment from 2010, asserted that there are ways to adjust their Facebook privacy settings to make their profile more secure (CBS, 2010). Facebook should have made the user profile secure by default, and the opportunity was given to users who decided to relax the privacy and security setting. The current Facebook privacy and security setting is Insecure by default. It is very complicated for users to change these settings to make it easier for Facebook to share user information with other websites. For example, Facebook can collect data on users’ “Like” buttons, which Facebook can later sell to third-party companies for targeted advising and profile profiling. The privacy feature and consent structure of Facebook by default, the user gets no say if they want their data released to other websites and companies. Facebook makes that decision for them. Unless they know how to shut it off, it will stay on permanently (FTC, 2019).

Interacting with people on Facebook has its accessibility rules. What may be visible for some people may not be visible for others, depending on the privacy settings of all the users involved in a post. In comment and owner interaction, a person’s comment may be visible beyond what they were expecting to reach. For instance, when a person changes their privacy for a post to the public instead of friends only, anyone who has commented on the post will have their interactions shown publicly without any notification explaining the change sent to them. Many people disagree with that as it shows their intended reach for their interaction was changed to beyond what they intended. Another issue is raised when a person’s interaction is not visible because of the past owner’s change of tag in their post. When person “A” creates a post and tags person “B,” with the privacy setting set to “Friends Only,” everyone in A’s and B’s friends list will be able to interact with the post. When “A” removes “B’s” tag, everyone who commented on the post from B’s list and who is not on A’s friends’ list will not be able to comment or like in that post. They will also not be able to access their comments or likes in that post and not even see it in their activity log. This raises an issue for those people who lost accessibility to the post as they cannot see their interactions while others can see them (Sayin et al., 2019).

3.2 Snapchat Case Study Analysis - Perceived Privacy

Since its inception, Snapchat has been used and recognized as one of the leading popular social media platforms available. However, its unique features have raised debate about the privacy rights of an individual and drawn distinctive criticism stemming from the company’s leadership and internal procedures. Similar to other Social Media applications, Snapchat has been the victim of security exploits. As a result, the company has had to publicly disclose the nature of these incidents and the corrective tactics utilized in remediating them. Although social media platform such as Snapchat has faced their fair share of data security and user privacy concerns, it remains very popular amongst Social Media platforms.

In 2013, Snapchat’s software fell victim to exploitation, in which 4.6 million users’ details were leaked (Krebs, 2014). In the months predating this breach, researchers at Gibson Security, an Australian firm, had notified Snapchat that their software was found to be vulnerable to an exploit in which an attacker could “trawl” for name and number combinations (Krebs, 2014). This vulnerability allows an attacker to take advantage of the “Friend Finder” function on the application. As a result, a user can look up a person’s phone number to find the associated user. In addition, rate-limiting was not an enabled function found on the application, thereby allowing a user to search thousands of records at a time. However, Snapchat did not take any meaningful action citing it as theoretical and not practical. As a result, Gibson Security publicly discloses the vulnerability, which within a week led to the publication of the “SnapchatDB.info” database online and forced Snapchat to disable the feature for their Friend Finder function on their application as well as implement appropriate rate-limiting (Krebs, 2014).

One of the issues prevalent amongst social media usage, especially Snapchat, is the user’s perception of privacy when using the service. This platform differentiates from others in that the “users control the visibility of the contents they share with others by defining how long these contents may be available” (Rauzzino and Correa, 2017). With this in mind, we must focus on the usage paired with this social media platform amongst the population. Specifically, this social media application has notably been known to be used for sending funny pictures and selfies, but, most sensitive of all, the platform has been used for sexting and sending sexual and pseudo-sexual material. A 2017 study (Rauzzino and Correa, 2017) aimed to identify the differences amongst both sexes and all socioeconomic status about the individually placed importance of their private lives. Since the platform is especially popular amongst millennials, the study utilized 268 users between 18 and 25, with sixty percent of the sample being female (Rauzzino and Correa, 2017). The study results stated that perceived privacy

was an important concern overall by the sampled individuals, with two exceptions. It was found that the female population placed more importance on users knowing the use of their personal information and on wishing they could be notified “every time a company looks into their personal use of Snapchat” (Rauzzino and Correa, 2017). Figure 2 below illustrates participants’ responses to the statement “I know about the use that is given to my info in my profile” (labeled D) (Rauzzino and Correa, 2017). This statement particularly highlights the truthfully skeptical mindset witnessed when asked if we know how Snapchat is using our data as well as other social media platforms.

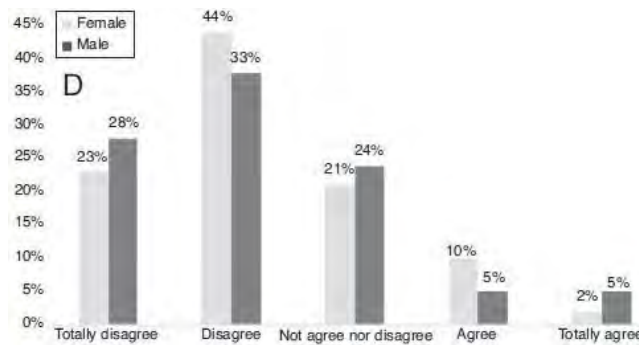


Figure 2: Participants Response to "I know about the use that is given to my info in my profile"(Rauzzino and Correa, 2017)

With this in mind, it is important for individuals who use the platform to be aware of their settings and manage them according to their privacy preferences. This issue is a combination of technical and social education. It does not seem to be disappearing any time soon, especially with the rise of Social Media usage in recent years stemming from the technological developments that have come about in the past decade.

3.3 Discussion

Several steps have been taken regarding online privacy and its concerns. Laws, regulations, and social education have been key components in achieving better practices concerning online privacy. Yet, based on the two case studies above, it is obvious that their privacy gaps range from poor security control implementation, education, and poor privacy legal control enforcement. Although some Social Media companies do a good job providing user-required privacy policies, there currently lack a shorter and clearer version of the most significant details of the policy so that consumers can know what to expect beforehand and determine whether to continue reading the entire policy or not (CBS, 2010). Based on the analysis of the two cases, the current security controls on most of these platforms are inadequate, and the gapping privacy and security holes on these platforms and related applications can lead to serious user data privacy violations.

4. EDEE Security Model

To address the issue with Social Media platform security, we present key elements needed in making best-case decisions to help users and organizations maintain important individual privacy while using social platforms. We present the Social Media EDEE model, as depicted in Figure 3.

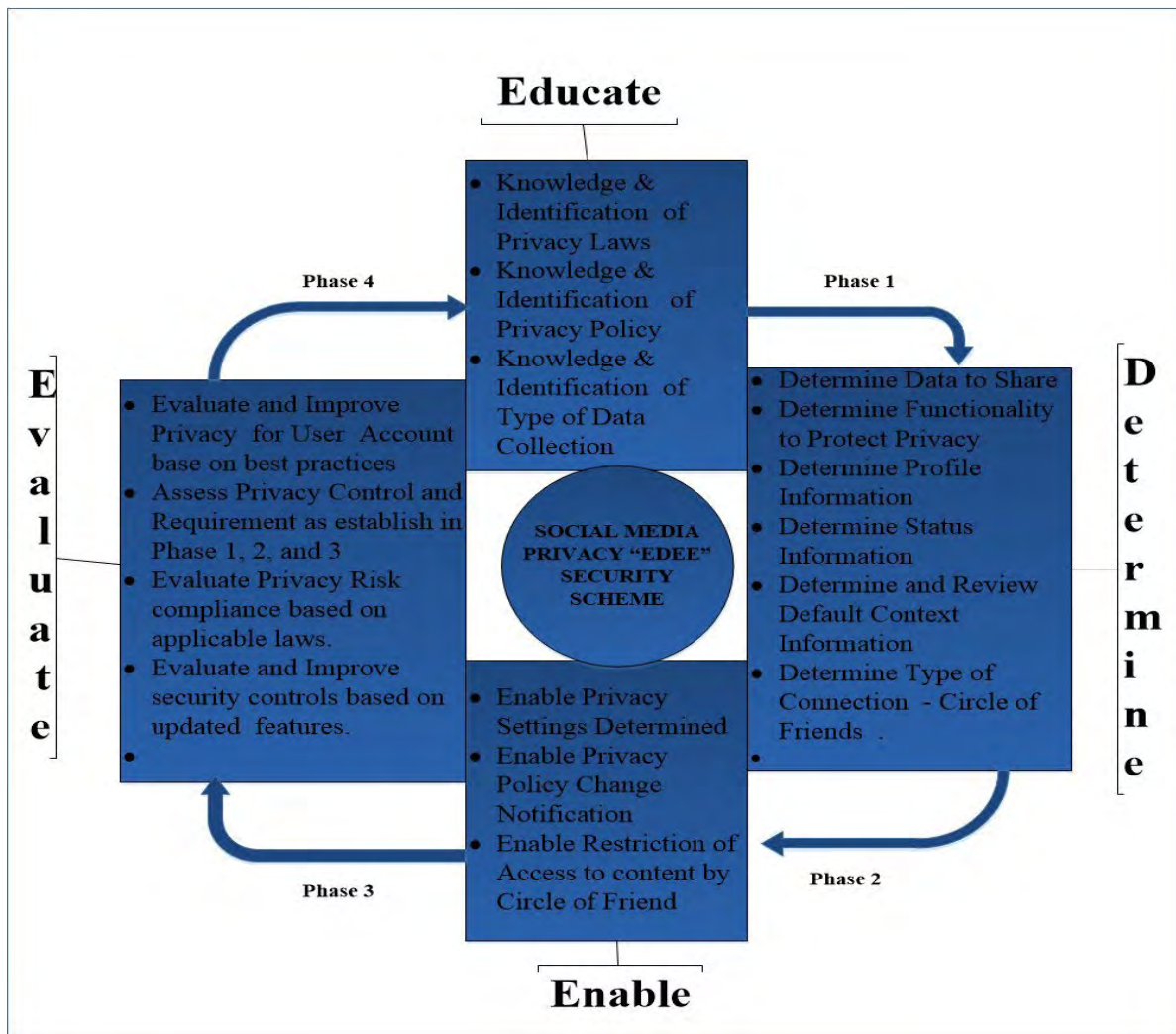


Figure 3: Social Media Privacy Through "EDEE" Security Model

As per the EDEE model, “; in Phase 1, the “Educate” Phase, the user gains knowledge about all applicable user privacy laws within the jurisdiction the platform is being used from; and the privacy policy of the platform you are using to determine what type of information the law protect and what the social media platform can collect on you. For example, suppose the user uses the platform from any European Union countries. In that case, the users will explore the data privacy rights protected under GDPR and the privacy policy of the specific platform they are using. In addition, at this phase, the users also understand the privacy implications of social media data breaches, resulting in reputation or financial loss.

In Phase 2, the “Determine” phase, the user determines what information to share and what functionality to use to protect privacy based on phase 1. This includes the user’s decision on profile information, including personally identifiable or demographic information. In determining profile information, the user’s role is to provide as minimum information as possible. Other information to consider here also includes the type of status update the user will post to identify an activity or event that an attacker can leverage to conduct a social engineering attack. This means users post very minimal information about present activities about themselves or the people they know. Also, in this determining phase for users to review, non-interactive information such as context data includes location information gathered and what to share with the social media platform. Last but not least is determining the content to share with and whom to share those content with. In this particular area, the idea is to have a sharing concept where users minimize privacy exposure by restricting sharing content to trusted friends within the user cycle.

In Phase 3, the “Enable” Phase, the user determines privacy settings on the social platform and enables privacy settings that disable public information sharing. The goal here is to understand the information shared by the Social Media platform publicly by default and the other information that a user shares yet can be turned off.

This includes and is not limited to a case where the platform can share information, and the user determines if the settings exist to enable who can access the content. The user can enable privacy policy change notifications. Any time privacy policy changes, the user can be aware of what has changed and how that impacts current user information or future content. The other consideration included determining approved contact such as friends and individual you follow and determining who see certain content. Further, the user enables options to disable the third application that, by default, gain access to user information.

In Phase 4, the Evaluate phase, the user actively assesses that the privacy controls enabled are functional through evaluations Phase 1, Phase 2, and Phase 3. These phases are cyclic as the user using this scheme will continue to Phase 1(Educate), Phase 2(Determine), Phase 3(Enable), and Phase 4(Evaluate). The EDEE Security Scheme allows users to observe respective privacy rights and treat them as property rights after the customer initiates the social media platform or application. Being mindful of online presence is helpful as sharing too much unnecessary information on social media can be shown to anyone and collect information. Therefore, users should only release information they feel comfortable sharing with others and not share unnecessary birth dates and locations.

5. Conclusions

Social Media platforms have been a very important part of many people's lives, considering the amount of PII and interactive. Other related information shared, the issue of privacy and security of such information should be addressed by the organization running these companies and while at the same time individuals made away of the privacy implications associated with the use of Social Media platform and effort that can be invested to protect the information they share. We demonstrated theoretically that based on the available information, some of these Social Media platforms continue to have vulnerabilities that can leave users subject to attack. Such information proves that the current security controls on most of these platforms are inadequate, and the gapping privacy and security holes on these platforms or related applications can lead to serious privacy and security risk. Our EDEE security model demonstrated that users need to be more educated about how the information they share could target them by advertisers if not to hack them. Users also need to be more aware of setting their privacy settings to achieve a perfect balance between using Social Media without sacrificing much personal information. Finally, in this work, we have demonstrated Social Media platform Security EDEE Security model can address the evolving Social Media platform user privacy concerns.

References

- CBS. (2010) "Facebook Privacy" [Video File]. Retrieved from https://www.youtube.com/watch?v=smF1ZV7vkw&feature=emb_title
- Center E. (2020) "EPIC - In re Facebook - Cambridge Analytica." Epic.org, Available:<https://epic.org/privacy/facebook/cambridge-analytica/>.
- Critchley, T. (2018) "The threat on the end of the phone: the danger of contact centre agents." Computer Fraud Security, 2018(2), 13–15.[https://doi.org/10.1016/S1361-3723\(18\)30015-0](https://doi.org/10.1016/S1361-3723(18)30015-0)
- FTC imposes restrictions on imposes \$5 billion Penalty and Sweeping New Privacy Restrictions on Facebook. (2019, July 24). Retrieved from <https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions>
- Heiligenstein M. X (2021) "Facebook Data Breaches: Full Timeline through 2022", Firewalltimes.com, 2022. [Online]. Available: <https://firewalltimes.com/facebook-data-breach-timeline/> [Accessed: 30- Jan- 2022].
- Huffman, B. (2015) "Developments in social media: First Amendment, Privacy, and misappropriation." Business Lawyer, 71(1), 305–319
- Isaak, J., Hanna, M. J. (2018) "User Data Privacy: Facebook, Cambridge An- Analytica, and Privacy Protection." The Policy Corner, 56–59. Retrieved from <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8436400>
- Jones D. (2020) "COMPANY NEWS; YOUTHSTREAM TO ACQUIRE SIXDEGREES FOR \$125 MILLION", Nytimes.com, 2020. [Online]. Available:<https://www.nytimes.com/1999/12/16/business/company-news-youthstream-to-acquire-sixdegrees-for-125-million.html>. [Accessed: 20- Apr- 2020].
- Kerner, S.M. (2018) "Facebook Data Breach Extended to Third-Party Applications." EWeek, 1-2. Retrieved from <http://search.ebscohost.com.libproxy.albany.edu/login.aspx?direct=true&db=a9h&AN=132117848&site=ehost-liv>
- Krebs, B (2014) "Target and Snapchat suffer major data breaches" . . Computer Fraud Security, 2014(1), 1,3–1,3. [https://doi.org/10.1016/S1361-3723\(14\)70001-6](https://doi.org/10.1016/S1361-3723(14)70001-6)
- Kryder, C. (2012) "Social Media and Privacy Issues: A Matter of Common Sense." AMWA Journal: American Medical Writers Association Journal, 27(1), 36. retrieved from <http://search.ebscohost.com.libproxy.albany.edu/login.aspx?direct=true&db=a9h&AN=85712790&site=ehost-live>
- Lamdan, S. S. (2015) "Social Media Privacy: A Rallying Cry to Librarians." Library Quarterly, 85(3), 261–277. <https://doi.org/10.1086/681610>

- McIntyre, K. (2014) "The Evolution of Social Media from 1969 to 2013: A Change in Competition and a Trend Toward Complementary, Niche Sites". *The Journal of Social Media in Society*, 3(2). Retrieved from <https://thejsms.org/tsmri/index.php/TSMRI/article/view/89/43>
- Nadeau, M. (2019) "General Data Protection Regulations (GDPR): What you need to know to stay compliant." Retrieved from <https://www.csoonline.com/article/3202771/general-data-protection-regulation-gdpr-requirements-deadlines-and-facts.html>
- Ortiz-Ospina, E. (2019) "The Rise of Social Media." Retrieved from <https://ourworldindata.org/rise-of-social-media>
- Perrin, A. (2015) "Social Media Usage: 2005-2015". Washington, D.C.: Pew Internet American Life Project. Retrieved October 12, 2015, from <http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/>
- PEW (2020) "Demographics of Social Media Users and Adoption in the United States," Pew Research Center: Internet, Science Tech, 2020. [Online]. Available:<https://www.pewresearch.org/internet/fact-sheet/social-media/>. [Accessed: 21-Apr-2020]
- Rauzzino, A., Correa, J. (2017) "Millennials' sex differences on Snapchat perceived privacy." *Suma Psicológica*, 24(2), 129–134.<https://doi.org/10.1016/j.sumpsi.2017.08.002>
- Roberts, M. (2019) "The Cyber Threat and Globalization: The Impact on U.S. National and International Security." By Jack A. Jarmon and Pano Yannakogeorgos. Lanham, MD: Rowman Littlefield, 2018. *Journal of Strategic Security*, 11(4), 85–88.<https://doi.org/10.5038/1944-0472.11.4.1716>
- SAYIN, B., ŞAHİN, S., KOGİAS, D. G., PATRIKAKIS, C. Z. (2019) "Privacy issues in post dissemination on Facebook." *Turkish Journal of Electrical Engineering Computer Sciences*, 27(5), 3417–3432.<https://doi.org/10.3906/elk-1811-25>
- Viala V. (2018) "Data Breaches Compromised 4.5 Billion Records in First Half of 2018*", Gemalto.com, 2020. [Online]. Available:<https://www.gemalto.com/press/pages/data-breaches-compromised-4-5-billion-records-in-first-half-of-2018.aspx>.
- Waldvogel, D. (2017) "Social Media and the DOD: benefits, risks, and mitigation. (COMMENTARY)(Report)". *Air Space Power Journal*, 31(2), 119–125.
- Whiting, R. (2018) "Don't Be Evil, Move Fast, Think Different: How Your Social Media Phone Applications Work against Your Privacy." *Thomas Jefferson Law Review*, 41(1), 127–162. Retrieved from <http://search.ebscohost.com.libproxy.albany.edu/login.aspx?direct=true&db=a9h&AN=134978314&site=ehost-live>
- Yankson, B., Iqbal, F., Aleem, S., Shah, B., Hung, P. C. K., and de Albuquerque, (2019) "A Privacy-Preserving Context Ontology (PPCO) for Smart Connected Toys," 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), Copenhagen, Denmark, 2019, pp. 1-6.
- Yankson, B., Iqbal, F., and Hung, (2021) "Systematic privacy impact assessment scheme for smart connected toys data privacy compliance" *Int. Journal of Big Data Intelligence* 8(1), , pp. 47-66.

PhD Research Papers

The Impact of CISO Appointment Announcements on the Market Value of Firms

Adrian Ford¹, Ameer Al-Nemrat¹, Seyed Ali Ghorashi¹ and Julia Davidson²

¹School of Architecture, Computing and Engineering, University of East London, UK

²Royal Docks School of Business and Law, University of East London, UK

a.ford1701@uel.ac.uk

Abstract: Previous studies concerning the economic impact of security events on publicly listed companies have focussed on the negative effect of data breaches and cyberattacks with a view to encouraging firms to improve their cyber security posture to avoid such incidents. This paper is an initial study on the impact of investment in human capital related to security, specifically appointments of chief information security officers (CISO), chief security officers (CSO) or similar overall head of security roles. Using event study techniques, a dataset of 37 CISO type appointment announcements spanning multiple world markets between 2012 and 2019 was analysed and statistically significant (at the 5% level) positive cumulative abnormal returns (CAR) of around 0.8% on average were observed over the three-day period before, during and after the announcement. Furthermore, this positive CAR was found to be highest, at nearly 1.8% on average, within the financial services sector and showing statistical significance at the 1% level. In addition to the industry sector, other characteristics were investigated such as job title, reporting structure, comparison of internal versus external appointments, gender and variations between markets. Although these findings were not as conclusive they are, nevertheless, good pointers for future research in this area. This overall positive market reaction to CISO related announcements is a strong case for publicly listed firms to be transparent in such appointments and to, perhaps, review where that function sits within their organisation to ensure it delivers the greatest benefits. As 24% of the firms analysed were listed outside the US, this study also begins to counter the strong US bias seen in similar and related studies. This research is expected to be of interest to business management, cyber security practitioners, investors and shareholders as well as researchers in cyber security or related fields.

Keywords: cyber security, CISO, CSO, market value, economic impact, event study

1. Introduction

The US Federal Bureau of Investigation (IC3 2020) reported a 69% increase in internet crime related complaints from 2019. With cybercrime so rife, along with damaging high-profile data breaches such as that of Marriott and British Airways (Ford et al. 2021a), cyber security should be very much a concern for organisations globally. Indeed, the UK Department for Digital Culture, Media and Sport report that 77% of businesses view cyber security as a high priority at board level (DCMS 2021) and Gartner (2021) predict that by 2025, 40% of boards will have cyber security committees established and overseen by a “*suitably qualified*” executive. Therefore, given the importance of cyber security, in the case of publicly listed companies one could reasonably expect markets to react in a positive way to news of investment in this area. This study is concerned with announcements of investment in human capital, specifically heads of security at executive level such as chief information security officer (CISO), chief security officer (CSO) or similar roles¹, the primary research question being what is the impact (if any) of this information on the market value of firms? The ability to measure and clearly identify any positive impact would surely encourage organisations to both invest in, as well as publicise establishment or improvement of their security functions and thereby their overall cyber security posture². Existing literature relating specifically to CISOs is rather sparse (Karanja & Rosso 2017) so this small scale initial study would also begin to fill a knowledge gap. Such research is expected to be of interest to business management, cyber security practitioners, investors and shareholders as well as researchers in cyber security or related fields.

2. Related work

Looking elsewhere, due to the dearth of literature on CISOs as reported by Karanja and Rosso (2017), Chatterjee, Richardson and Zmud (2001) examined the economic impact of the creation of new chief information officer (CIO) positions using event study techniques and observed positive market reactions of 1.16% on average and even higher (almost 3%) for heavily information technology (IT) driven industry sectors. This variation between sectors was also reported by Tweneboah-Kodua, Atsu and Buchanan (2018) who observed that financial services firms reacted more rapidly and more significantly than those in the technology sector. Their study focussed on

¹ For convenience, hereinafter the acronym ‘CISO’ will be used as a generic term to refer to roles of this type.

² Cyber security posture is a term encompassing training and awareness in addition to governance and technical solutions.

the negative impact of cyberattacks on stock returns and a systematic literature review by Spanos and Angelis (2016) found that 76% of studies in this area reported an impact which was both negative and statistically significant. For instance, Lin et al. (2020) observed losses of 1.44% on average over a five-day window. Andoh-Baidoo, Amoako-Gyampah and Osei-Bryson (2010) found -3.18% abnormal returns over a three-day period. Cavusoglu, Mishra and Raghunathan (2004) observe -2.1% on average over the two days following the announcement. Goel and Shawky (2009) cite -1% in the days immediately surrounding the event.

Moving back to positive economic impacts, literature in this area was found to be lacking. Cavusoglu et al. (2004) found a positive impact on security developers after the disclosure of security breaches by affected firms. Deane et al. (2019) studied ISO 27001 certification announcements and found the abnormal market returns both positive and statistically significant (0.72% on average over two days). A more recent systematic literature review (Ali et al. 2021) updating the work of Spanos and Angelis (2016) acknowledges the relative lack of studies concerning the positive impact of favourable information security events such as regulation, certification or investment also noting that such favourable events generate lower magnitude abnormal returns (in the range 0.63% to 1.36%) than unfavourable events which could lose up to 10% of market value.

This study, therefore, will not only increase the CISO research knowledgebase but also help address the shortfall in studies on favourable security events. Ali et al. (2021) also note that 76% of existing similar studies were based solely on US data, so the aim here was also to look globally as far as possible (given English language restrictions) in an attempt to offset this US bias (cf. Ford et al. 2021b).

3. Methodology

From a high-level perspective, the approach to this research was first, to hand gather a dataset of CISO appointment announcements, then filter this dataset keeping only those examples related to publicly listed companies. The final step was to identify any impact of these announcements on the share price of each firm through event study techniques.

3.1 Event studies

Event studies are a widely reported method to assess the impact of a specific event on the share price (market value) of firms - a detailed description may be found in e.g. MacKinlay (1997). A key assumption of this method is the ability of the market to reflect all available information as described by the efficient market hypothesis (i.a. Fama 1970). Through the observation of share price movements in reaction to information regarding a specific event, such as an executive appointment announcement, over a short time period (the event window), it is possible to ascertain how the market reacted to that specific event, with the caveat of no other confounding events occurring during this period.

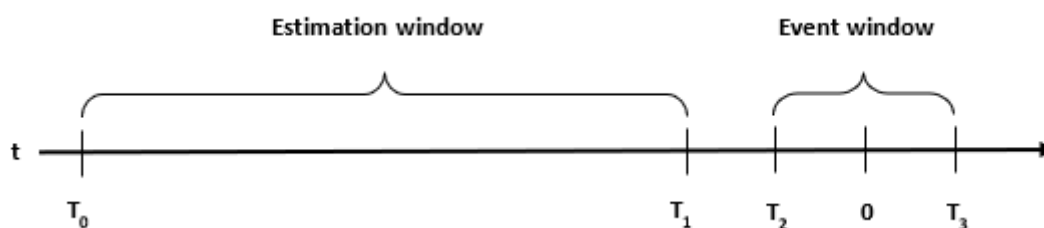


Figure 1: Event study timeline

A frequently used approach for information security related event studies is the market model (e.g. Cavusoglu et al. 2004; Andoh-Baidoo et al. 2010; Hinz et al. 2015; Schatz & Bashroush 2016; Castillo & Falzon 2018; Tweneboah-Kodua et al. 2018; Jeong et al. 2019; Ford et al. 2021a,b) which, through a regression analysis of a firm's share price over a relatively long estimation window (see **Figure 1**), is able to predict stock returns during the later (usually much shorter) event window. This approach assumes that returns follow a single factor model (1) where the return of firm i on day t ($R_{i,t}$) is dependent on the corresponding daily return of the reference market ($R_{m,t}$) and the extent of the security's responsiveness (β_i) offset by its abnormal return (α_i). The error term $\varepsilon_{i,t}$ is expected to be zero with finite variance. Abnormal returns are calculated for the event window (2) and reported as a cumulative abnormal return (CAR) over the whole event window (3). For cross-sectional analyses, a cumulative average abnormal return (CAAR) was calculated over N events (4).

$$R_{i,t} = \alpha_i + \beta_i \cdot R_{m,t} + \varepsilon_{i,t} \quad (1)$$

$$AR_{i,t} = R_{i,t} - (\alpha_i + \beta_i \cdot R_{m,t}) \quad (2)$$

$$CAR_i = \sum_{t=T_2}^{T_3} AR_{i,t} \quad (3)$$

$$CAAR = \frac{1}{N} \sum_{i=1}^N CAR_i \quad (4)$$

3.2 Data collection

The base dataset was hand-gathered from internet searches for CISO appointments³ building up a Microsoft Excel spreadsheet. Many results were clearly governmental organisations or not-for-profits or private companies, so were discarded. It was decided to cap the date range at 31/12/2019 in order to avoid market uncertainties due to COVID-19, that in itself being a long-term confounding event. Once the residual set seemed to be mostly publicly listed companies each data record was carefully reviewed, ensuring that the stock was still listed and share price data available. Data fields extracted included company name, announcement date, job title, position in the organisation (reporting line) and gender. Some announcements made it clear this was a newly created CISO position and so these were marked as such whereas others merely implied this – these were flagged separately due to the lesser information salience (e.g. Ramos, Latoeiro & Veiga 2020). Other records had to be filtered out as they were not listed before the start of the estimation window. A few examples (Bridgestone Americas, Santander UK and CareerBuilder) were related to subsidiary companies but these were kept as there would be a contribution, at least, to any market reaction despite less salience, also to maintain as large a dataset as possible. Once the filtering was complete, a dataset of 41 records remained. The final step was to look for confounding events. The dates of the nearest events (before/after day zero) were also recorded so these could be filtered out dynamically as needed. Confounding events were considered to be earnings/dividend announcements or another executive appointment (consistent with e.g. Chatterjee et al. 2001). There were three examples (Digital Realty, Wells Fargo, Axon) where joint announcements were made with other positions and these were included, again to maximise the dataset despite the lesser salience.

The importance of the choice of market index to use as a reference has been highlighted by e.g. Kannan, Rees and Sridhar (2007). Consistent with Ford et al. (2021b), the most appropriate market index was chosen in each case being, ideally, one in which the stock in question was included and favouring narrower, more focussed indices rather than a one-size-fits-all approach given, of course, that this index data was available from Yahoo!Finance (otherwise the next best was selected). Some firms had multiple listings, in which case the primary listing and associated index were used.

3.3 Data analysis

To facilitate the analyses, R (R Core Team 2018)⁴ code was developed to extract share price and index data directly from Yahoo!Finance for each data record. Event studies were then carried out through an R package (Schimmer, Levchenko & Müller 2014)⁵ applying the market model as described above. Announcements falling on non-trading days were defaulted to the next available trading day. An estimation window of 120 days was chosen, consistent with previous studies (e.g. Goel & Shawky 2009; Andoh-Baidoo et al. 2010; Schatz & Bashroush 2016; Richardson et al. 2019) ending one trading day before the event window in all cases. Tweneboah-Kodua et al. (2018) recommend avoiding overlap of the estimation and event windows in this way to avoid “*parameter contamination*”. Although the event window should be broad enough to contain any uncertainty in the date of the event, the longer the window, the less likely it is to detect abnormal returns (Dyckman, Philbrick, & Stephan 1984). Some studies observed a market reaction in advance of the event date due to information leakage, such as Lin et al. (2020) who found significant evidence of opportunistic pre-official announcement insider trading related to data breaches using event studies. For this exercise, a range of event windows were initially chosen starting from up to two days before the event and varying in length from 2 up to 20 trading days to catch any pre-event effects as well as others, for instance, sector specific effects reported by

³ Sources included ProQuest (CSO Online), Google and Bing

⁴ R version 4.1.2 (2021-11-01)

⁵ EventStudy package version 0.36.900 (API version 1.059)

i.a. Tweneboah-Kodua et al. (2018) who observed more rapid responses to information security events from e.g. the financial services sector thereby justifying longer event windows for sectorial comparison purposes.

3.4 Hypothesis development

The null hypothesis for event studies maintains that there are no abnormal returns within the event window. The standard deviation of abnormal returns during the event window is described by (5) where M_i refers to the number of non-missing returns. The t-value for the CAR over the event window for each firm i was then calculated according to (6).

$$S_{AR_i} = \sqrt{\frac{1}{M_i - 2} \sum_{t=T_0}^{T_1} (AR_{i,t})^2} \quad (5)$$

$$t_{CAR_i} = \frac{CAR_i}{\sqrt{(T_3 - T_2 + 1) S_{AR_i}^2}} \quad (6)$$

For cross-sectional analyses the t-statistic (t_{CAAR}) was calculated based on the CAAR (8) with S_{CAAR} being the standard deviation of the CARs for each firm i across the sample of size N (7).

$$S_{CAAR} = \sqrt{\frac{1}{N - 1} \sum_{i=1}^N (CAR_i - CAAR)^2} \quad (7)$$

$$t_{CAAR} = \sqrt{N} \frac{CAAR}{S_{CAAR}} \quad (8)$$

Significance testing in this way is consistent with e.g. Castillo and Falzon (2018), Deane et al. (2019) and Jeong et al. (2019). Indeed, Deane et al. (2019: 115) argue that “the t test is considered to be the best framework for analyzing statistical significance in most event study frameworks and to be relatively robust”.

4. Results and discussion

Event studies were carried out as above for 10 event windows of differing length across all 41 CISO appointment announcements. A visualisation of the overall results showed that event windows (-1, 1) and (-1, 0) were of interest and thus records with confounding events ± 2 days were filtered out⁶ leaving 37 records. The revised visualisation is shown in **Figure 2**. The most positive impact is seen around the three or two-day event windows (-1, 1), (-1, 0) with the market value reverting back to normal by day 4. There was also a slight peak at day 5, but of course, this may be due to confounding events. By day 20 it can be seen that Fortinet is the best performing stock with a CAR of almost 20%, and MGIC the worst.

⁶ This is a quite stringent approach, similar studies have only filtered out confounding events within the event window, this approach leaves a gap of one day in case of any advance information leakage of the confounding event.

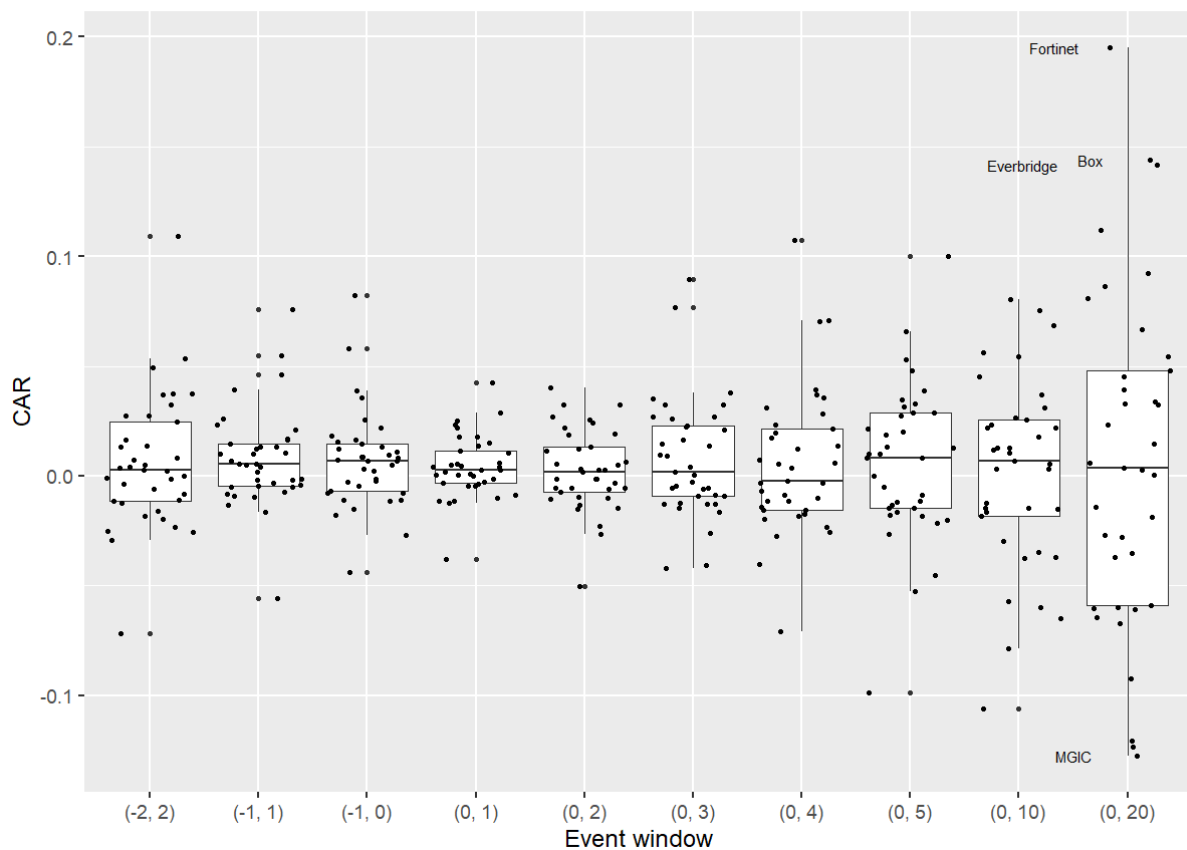


Figure 2: Comparison of event windows

A CAAR was calculated for multiple firms across each window and the results shown in **Table 1**. The three-day event window $(-1, 1)$ showed the most positive CAAR of almost 0.8% (significant at the 5% level), similar to the findings of e.g. Chatterjee et al. (2001) for CIO appointments with their dataset being of a comparable size as well (1.16% over 96 announcements). The fact that this includes a pre-event day indicates a little information leakage prior, as the date accuracy was carefully verified (cf. Lin et al. 2020). Within this window, 22 of the 37 events (59%) had abnormal returns of greater than zero, therefore this window was chosen as the basis for further analyses. Usage of this event window $(-1, 1)$ has been previously reported in studies of this type (e.g. Chatterjee et al. 2001; Andoh-Baidoo et al. 2010; Bose & Leung 2014; Khansa 2015; Modi et al. 2015). A lack of statistical significance for the longer windows is also seen, indicative of a tendency of market recovery towards zero abnormal returns over time as reported by Dyckman et al. (1984).

Table 1: CAAR by event window

Event Window	N	CAAR	t_{CAAR}	% Positive CAR
$(-2, 2)$	37	0.0062	1.2356	54
$(-1, 1)$	37	0.0077	2.1091**	59
$(-1, 0)$	37	0.0070	1.8997*	62
$(0, 1)$	37	0.0043	1.8334*	65
$(0, 2)$	37	0.0025	0.8411	51
$(0, 3)$	37	0.0075	1.6717	54
$(0, 4)$	37	0.0055	1.0091	49
$(0, 5)$	37	0.0059	1.0054	57
$(0, 10)$	37	0.0015	0.2074	59
$(0, 20)$	37	0.0070	0.5529	57
	370	0.0055		57

*, **, *** Represent statistical significance at the 10%, 5% and 1% levels respectively.

A visualisation of how the data changes over time is shown in **Figure 3**. Out of the 37 total events in the dataset, by far the majority (92%) occurred between 2017 and 2019 but the overall number of CISO announcements does

not seem to be increasing year-on-year - it actually dropped from 18 to 5 in the last two years sampled, so there seems to be a lack of awareness of the potential benefits of sharing this information with the market. A relative measure of the CAAR is also shown flipping from positive in 2012 to negative in 2015 and remaining positive for the bulk of the dataset in the final three years.

Regarding the origin of the appointment, internal appointments were low (27%) in 2017 dropping to 11% in 2018 and subsequently to zero in 2019. So, if we ignore the sparse data in previous years we have a trend toward solely external recruitment. Interestingly the internal appointments, although less at 6 out of 37 (16%), generated a CAAR of 1.28% which was almost double that of the way more frequent external appointments (CAAR=0.67%) although these results were not statistically significant. Chatterjee et al. (2001) suggest that the market may respond better to internal appointments because the appointee would 'hit the ground running' due to existing in-depth knowledge of the business and established relationships with management.

Karanja and Rosso (2017) observed in their dataset of 55 CISOs spanning 2010-2014 that only 11% were female. Although here it is difficult to identify any clear trend over time, it can be noted there were no such examples in the data sample prior to 2017 (although only three data points) yet in the final three years the percentage of women was fairly constant (18, 22, 20%) so despite the lack of any obvious trend in the data here, comparing with Karanja and Rosso (2017) we see almost double the percentage of women CISOs – a refreshing increase in diversity. That said, the 30 examples (82%) of male appointees yielded a CAAR of +0.85% significant at the 5% level whereas the remaining 7 female appointments generated only +0.38% which was, unfortunately, not statistically significant.

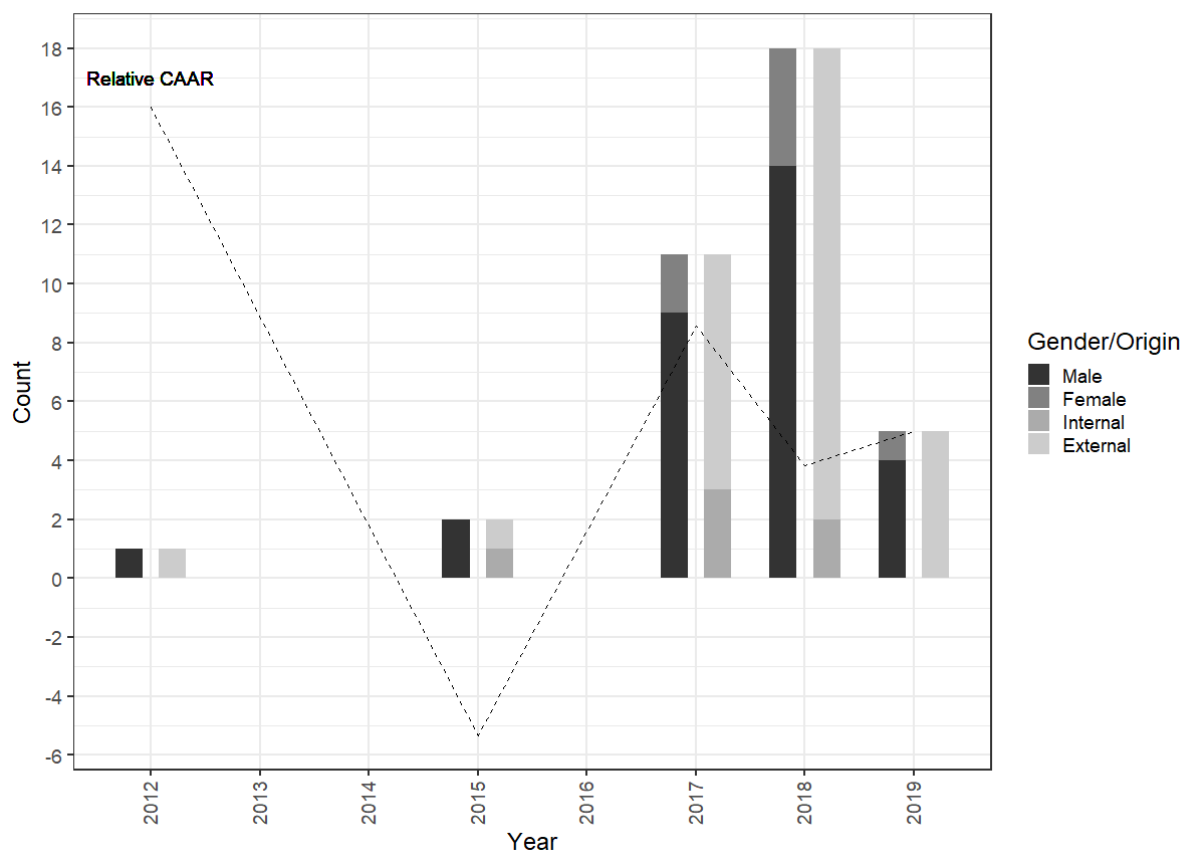


Figure 3: Breakdown by year

A breakdown of CAAR by sector is shown in **Table 3**. The financial services sector was the clear winner showing positive CARs of almost 1.8% on average and statistical significance at the 1% level. Such sector specific behaviour has been highlighted in e.g. Tweneboah-Kodua et al. (2018). It is interesting to note that 17 out of 37 announcements sampled (46%) belonged to this sector perhaps indicating more willingness (or regulation, of course) for transparency within this industry. It does seem surprising that negative market sentiment was

identified for 10 examples spanning 4 sectors especially one as sensitive as communication services which came in at more than 1% negative. Nevertheless, there is a deficiency of data here and no statistical significance.

Table 2: CAAR by industry sector

Industry Sector	N	CAAR	t _{CAAR}	% Positive CAR
Financial Services	17	0.0178	3.3027***	76
Healthcare	1	0.0169		100
Consumer Cyclical	2	0.0077	3.2553	100
Utilities	1	0.0066		100
Technology	6	0.0046	0.5974	50
Real Estate	1	-0.0047		0
Consumer Defensive	1	-0.0082		0
Industrials	5	-0.0084	-0.6592	40
Communication Services	3	-0.0101	-2.8456	0
	37	0.0077		59

*, **, *** Represent statistical significance at the 10%, 5% and 1% levels respectively.

Looking more closely at the content of the announcements, information on job title and position in the organisation are shown in **Table 4** and **Table 5** respectively. The most positive returns seem to occur when the CISO title is combined with a VP/SVP position – clearly this is recognised by the market as having more influence within the organisation although due to lack of data it is not statistically significant. The title CSO (a broader role not restricted to only information security) does show some significance but only at the 10% level. Even the single example of appointment of a Deputy CISO role (occurring shortly after the CISO was recruited) showed positive CAR at nearly 0.5%. Interestingly, combination of the security role with the CIO title results in the least positive CAR, or with that of trust (CTSO) yields negative of 1.7%. It appears that the market is preferring a focussed role with a high level of influence.

Table 3: CAAR by job title

Title	N	CAAR	t _{CAAR}	% Positive CAR
VP/CISO	3	0.0358	2.4170	100
SVP/CISO	1	0.0234		100
CSO	6	0.0098	2.0353*	83
Global Head of Cyber Risk	1	0.0053		100
Deputy CISO	1	0.0048		100
CISO	23	0.0043	0.8788	43
CIO/CISO	1	0.0017		100
CTSO	1	-0.0165		0
	37	0.0077		59

*, **, *** Represent statistical significance at the 10%, 5% and 1% levels respectively.

Carrying forward the argument of influence we can get some idea of the position in the organisation from some (16 out of 37 were not specified)⁷ of the announcements where the reporting line was cited in **Table 5**. First of all, the results where data were available were not significant, so the null hypothesis cannot be rejected, although it is interesting to note that the expected result of direct CEO reporting (and therefore greatest influence) being the highest CAAR is not the case - it seems that COO or CFO reporting is more well received by the market. If the concept of operations is combined with technology (CTO) or EVP though, the result is negative! The observation that a reporting line into the CIO yields negative CAAR seems consistent with Williams (2007) and the conflict of interest argument – would there be reluctance on the part of such a CISO to call out security flaws in the CIO's IT environment?

Table 4: CAAR by CISO reporting line

Reporting line	N	CAAR	t _{CAAR}	% Positive CAR
COO	2	0.0215	0.8737	50
CFO	1	0.0162		100
CEO	4	0.0148	1.3565	75
Corporate Secretary	1	0.0134		100
CTO	1	0.0130		100

⁷ In some cases, it could be assumed from the announcement, however unless explicit it was classified as unspecified.

Reporting line	N	CAAR	t _{CAAR}	% Positive CAR
(not specified)	16	0.0107	1.7592*	62
CISO	1	0.0048		100
CTO/COO	1	-0.0013		0
CIO	8	-0.0033	-0.3978	50
EVP Operations	1	-0.0047		0
SVP	1	-0.0050		0
	37	0.0077		59

*,**,*** Represent statistical significance at the 10%, 5% and 1% levels respectively.

The Chatterjee et al. (2001) study was restricted to the establishment of newly created CIO positions only. If the CISO announcements here contained a clear or implied indication that this was a new CISO position this was captured in the data record and analysed as shown in **Table 6**. The results were almost the exact opposite of expected, that the market would react better if an organisation was currently lacking an established security function. Perhaps the market expectation is that there should already be such a function in existence? The results are not significant except those cases where new or established was not specified which was a clear positive CAAR approaching 1.2%.

Table 5: Analysis of new or established CISO roles

New or established?	N	CAAR	t _{CAAR}	% Positive CAR
(not specified)	25	0.0115	2.6205**	64
New (implied)	4	0.0015	0.2505	50
New (specified)	8	-0.0013	-0.1501	50
	37	0.0077		59

*,**,*** Represent statistical significance at the 10%, 5% and 1% levels respectively.

In view of the intended international nature of this study, an analysis by different markets is shown in **Table 7** using market currency as a primary key (to be less granular due to data deficiency). Here India shows by far the highest CAR of over 7.6% but with only one example. It is the UK and US who show significantly positive CAARs at the 5% level (with the UK almost double that of the US). Only Israel and Australia are displaying negative CAARs, but again lacking data. Despite no intended US focus, we see that 28 out of 37 (76%) announcements originated from US markets so there is little opportunity to analyse other markets in detail. This percentage matches closely the observation of Ali et al. (2021) on US dominance in such research who attribute this to regulatory effects encouraging transparency in addition to language restrictions. Over time there is an expectation that transparency (and thus availability of data) will increase in other markets as new regulations come into force.

Table 6: Analysis by market currency

Market Currency	N	CAAR	t _{CAAR}	% Positive CAR
INR	1	0.0757		100
GBP	2	0.0158	14.3636**	100
USD	28	0.0081	2.4346**	61
JPY	1	0.0053		100
EUR	2	0.0052	0.6242	50
ILS	1	-0.0013		0
AUD	2	-0.0320	-1.3453	0
	37	0.0077		59

*,**,*** Represent statistical significance at the 10%, 5% and 1% levels respectively.

5. Conclusion

It has been shown in this introductory study that the announcement by a publicly listed company of a CISO appointment does indeed induce a positive market reaction, particularly within the financial services sector (+1.8% in the three days surrounding the event). In the data sample of 37 announcements analysed, 17 were financial services companies indicating a sector specific willingness to report, or regulatory effect.

Based on our analyses, we would advise the following content for a CISO appointment announcement to deliver maximum positive abnormal market return (ideally the hiring organisation would be a US or UK listed company in the financial services sector): job title CISO with VP or SVP responsibility stipulated (no mention whether this

is a new or established role), internal placement, male, reporting to the COO, CFO, or CEO but definitely not the CIO!

Chatterjee et al. (2001) calculated an approximate range of US\$7.5m (median approach) to US\$76m (mean approach) increase in market capitalisation for CIO appointments thereby easily justifying “*the trend in escalated executive salaries*”! Repeating this estimation across this whole CISO dataset gives a range of US\$94m to US\$318m, so it seems that this trend is very much continuing and clearly applies to CISOs as well!

This research should highlight the clear economic benefit of CISO appointments, as well the advantage of transparency in this area, to business management through the value the market places on the CISO role. There is an opportunity for other sectors to follow the lead of financial services and for other markets to adopt US practices and get ahead of the curve before new regulations come into effect. Once more data becomes available this exercise could be revisited for more in-depth analyses.

Acknowledgements

The authors wish to thank the anonymous reviewer for their helpful comments in improving this manuscript.

References

- Ali, S.E.A., Lai, F.-W., Dominic, P.D.D., Brown, N., Lowry, P.B., Ali, R.F. (2021), *Stock market reactions to favorable and unfavorable information security events: A systematic literature review*, Computers & Security, **110**, 102451
- Andoh-Baidoo F.K., Amoako-Gyampah K., Osei-Bryson K.M. (2010), *How Internet security breaches harm market value*, IEEE Security and Privacy **8**(1), 36–42
- Bose, I., Leung, A.C.M. (2014), *Do phishing alerts impact global corporations? A firm value analysis*, Decision Support Systems, **64**, 67–78
- Cavusoglu, H., Mishra, B., Raghunathan, S. (2004), *The Effect of Internet Security Breach Announcements on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers*, International Journal of Electronic Commerce, **9**, 69–104
- Chatterjee, D., Richardson V.J., Zmud, R.W. (2001), *Examining the Shareholder Wealth Effects of Announcements of Newly Created CIO positions*, MIS Quarterly, **25**(1), 43–70, <https://www.jstor.org/stable/3250958>
- DCMS (2021), *Cyber Security Breaches Survey 2021*, <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2021/cyber-security-breaches-survey-2021>
- Deane J.K., Goldberg D.M., Rakes T.R., Rees L.P. (2019), *The effect of information security certification announcements on the market value of the firm*, Information Technology & Management, **20**(3), 107–121
- Dyckman, T., Philbrick, D., Stephan, J. (1984), *A Comparison of Event Study Methodologies Using Daily Stock Returns: A Simulation Approach*, Journal of Accounting Research, **22**, (Supplement)
- Fama, E. F. (1970), *Efficient Capital Markets: A Review of Theory and Empirical Work*, The Journal of Finance, **25**(2), 383–417
- Ford, A., Al-Nemrat, A., Ghorashi, S., Davidson, J. (2021a), *The Impact of GDPR Infringement Fines on the Market Value of Firms*, Proceedings of the 20th European Conference on Cyber Warfare and Security, <https://doi.org/10.34190/EWS.21.088>
- Ford, A., Al-Nemrat, A., Ghorashi, S., Davidson, J. (2021b), *The Impact of Data Breach Announcements on Company Value in European Markets*, WEIS 2021: The 20th Annual Workshop on the Economics of Information Security, <https://weis2021.econinfosec.org/wp-content/uploads/sites/9/2021/06/weis21-ford.pdf>. Accessed on: 21/12/2021
- Gartner Inc. (2021), *Gartner Predicts 40% of Boards Will Have a Dedicated Cybersecurity Committee by 2025*, <https://www.gartner.com/en/newsroom/press-releases/2021-01-28-gartner-predicts-40--of-boards-will-have-a-dedicated->
- Goel, S., Shawky, H.A. (2009), *Estimating the market impact of security breach announcements on firm values*, Information & Management, **46**(7), 404–410
- Hinz, O., Nofer, M., Schiereck, D., Trillig, J. (2015) *The influence of data theft on the share prices and systematic risk of consumer electronics companies*, Information & Management, **52**(3), 337–347
- IC3 - Internet Crime Complaint Center (2020), *Internet Crime Report 2020*, https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf
- Jeong, C., Lee, S., Lim, J. (2019), *Information security breaches and IT security investments: Impacts on competitors*, Information & Management, **56**(5), 681–695
- Kannan, K., Rees, J., Sridhar, S., (2007), *Market Reactions to Information Security Breach Announcements: An Empirical Analysis*, International Journal of Electronic Commerce, 01 September **12**(1), 69–91
- Lin, Z., Sapp, T.R., Ulmer, J.R., Parsa, R. (2020) *Insider trading ahead of cyber breach announcements*, Journal of Financial Markets, **50**, 100527
- Karanja, E., Rosso, M.A. (2017), *The Chief Information Security Officer: An Exploratory Study*, Journal of International Technology and Information Management, **26**(2), 23–47

- Khansa L. (2015), *M&As and market value creation in the information security industry*, *Journal of Economics and Business*, **82**, 113–134
- Ramos S.B., Latoeiro P., Veiga, H. (2020), *Limited attention, salience of information and stock market activity*, *Economic Modelling*, **87**, 92-108
- Mackinlay, A. C. (1997), *Event Studies in Economics and Finance*, *Journal of Economic Literature* **35**(1) (March)
- Modi, S.B., Wiles, M.A., Mishra, S. (2015), *Shareholder value implications of service failures in triads: The case of customer information security breaches*, *Journal of Operations Management*, **35**, 21–39
- R Core Team (2018), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available on: <https://www.R-project.org/>
- Richardson, V.J., Smith, R.E, Watson, M.W. (2019) *Much Ado about Nothing: The (Lack of) Economic Impact of Data Privacy Breaches*, *Journal of Information Systems*: **33**(3), 227-265
- Schatz, D., Bashroush, R. (2016), *The impact of repeated data breach events on organisations' market value*, *Information & Computer Security*, **24**(1), 73-92
- Schimmer, M., Levchenko, A., and Müller, S. (2014), *EventStudyTools (Research Apps)*, *St.Gallen*. Available on: <http://www.eventstudytools.com>. Accessed on: 26/02/2021
- Spanos, G., Angelis, L. (2016), *The impact of information security events to the stock market: A systematic literature review*, *Computers and Security*, **58**, 216-229
- Tweneboah-Kodua, S., Atsu, F. and Buchanan, W. (2018), *Impact of cyberattacks on stock performance: a comparative study*, *Information and Computer Security*, **26**(5), 637-652
- Williams, P. (2007), *Executive and board roles in information security*, *Network Security*, **8**, 11-14
- Yahoo!Finance (2019), *Historical Data*, <https://finance.yahoo.com/quote>

What is a Substantial Contribution to a Research Project in Offensive Cyberspace Operations that Merits Co-Authorship?

Gazmend Huskaj

Department of Military Studies, Swedish Defence University, Stockholm, Sweden

School of Informatics, University of Skövde, Sweden

gazmend.huskaj@fhs.se

Abstract: This article reviews the question what is a substantial contribution to a research project in offensive cyberspace operations that merits co-authorship? Frustrations and conflicts may develop during research projects when researchers with different backgrounds, cultures, research fields and expertise decide to conduct research and produce and publish those results. The focus of this paper is a research project in cyberspace operations while taking into account the power-dynamics inherent in the academic system and how these can affect the co-authorship of research products. First, the purpose with doing research is presented. Next, three models of the research process are reviewed, describing their differences and similarities. Then, linguistic analysis is applied on a set of terms in guidelines for co-authorship described in some different universities in Sweden. The results present a model for a research project in offensive cyberspace operations and based on the output of the linguistic analysis, the model is used to quantify and describe what a substantial contribution is in three scenarios that merits co-authorship.

Keywords: Co-authorship, Linguistic Analysis, Offensive Cyberspace Operations, Power-dynamics, Substantial Contribution

1. Introduction

Frustrations and conflicts may develop during research projects when researchers with different backgrounds, cultures, research fields and expertise decide to conduct research and produce and publish those results (NTNU, 2021). The same frustrations and conflicts may also arise between a PhD-student and his/her supervisor(s), especially during a PhD research project that tends to span over several years. It is known that a PhD research project has psychological, emotional and other aspects like “enthusiasm, isolation, boredom, frustration, anxiety and euphoria” (Morrison-Saunders et al., 2010, p. 3) that can affect the PhD-student and his/her supervisor(s) in various ways. Frustrations have been reported between students and supervisors in cases of lack of feedback, supervisors changing their mind often, and students having to wait “for what they considered unreasonable periods, for supervisors to provide feedback and guidance on thesis drafts” (Morrison-Saunders et al., 2010, p. 9). Additional frustrations and conflicts have arisen in cases where supervisors have not been invested in the PhD-project, the supervisee felt lack of support, the supervisee being more experienced than the supervisor, and the supervisor being too busy (Nelson & Friedlander, 2001).

Another area where frustrations and conflicts may arise is during research projects where the PhD-student and his/her supervisor(s) co-author research products. While co-authoring affects how productive a researcher is, there are also addition positive aspects such as feeling safer (Tschannen-Moran & Nestor Baker, 2004), having more fun and to the actual production of a research product (Nygaard, 2017). The negative aspects can arise when the PhD-student and the supervisor(s) disagree on what a substantial contribution to a research project is because of differing perspectives the supervisee and supervisor(s) may have on co-authorship. In one co-authoring case, a supervisor took credit for a PhD-student’s work: “I think I’ve seen some kind of dark, underbelly of academia. ... But, when you’ve seen that a couple of times, you’re also going to start to get more wary. ... I think it’s a rotten, rotten thing, co-authorship” (Nygaard, 2017, p. 526).

A PhD-project spanning over several years, the psychological and emotional aspects coupled with supervisors that take time to provide feedback, not being invested in the PhD-project, not providing enough of support, being too busy, having different perspectives on co-authorship, and being inexperienced in comparison to the PhD-student, are all elements that lead only in one direction: frustration, conflict, and in some cases, the PhD-student being the one to take the fall (Times Higher Education, 2017). Therefore, it is important to review the question *what is a substantial contribution to a research project in cyberspace operations that merits co-authorship*.

The remainder of the paper is as follows: First, the reason why and those that do research is covered. Next, the research project is described followed by the method and linguistic analysis (Goldkuhl, 2002), which is applied throughout the paper. Finally, substantial is quantified followed by conclusions and why this is important.

2. Why do Research?

There are many reasons to why people do research:

“add to the body of knowledge, to solve a problem, to find out what happens, to find the evidence to inform practice, to develop a greater understanding of people and their world, to predict, plan and control, to contribute to other people’s well-being, to contribute to personal needs, to test or disprove a theory, to come up with a better way, to understand another person’s point of view, [or] to create more interest in the researcher” (Oates, 2005, pp. 17-21).

The outcomes of research are different research products (Oates, 2005). A reader with a background in the security and intelligence services will note the word “products.” Why? Because Intelligence is about knowing other States’ secrets and using those secrets to achieve one own’s national interests, but also to secure those national interests. Intelligence operations are also vital in supporting military operations (Joint Chiefs of Staff [JCS], 2013). The Intelligence process is about the actions as mentioned in Figure 1. These actions are done in a cycle, also known as the Intelligence Cycle. Once the analysis is done, a report is produced, and this report is known in Intelligence parlance as an “Intelligence Product.” Therefore, when Oates (2005) states that the outcome of research are different research “products”, this indicates that research may actually have a lot in common with Intelligence operations and the Intelligence process.

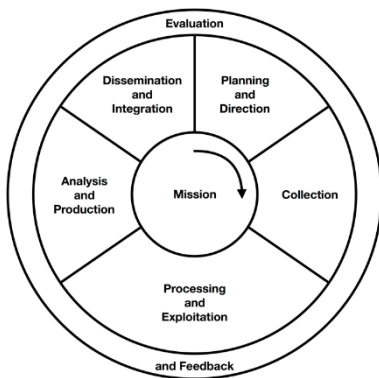


Figure 1: The Intelligence Cycle (JCS, 2013)

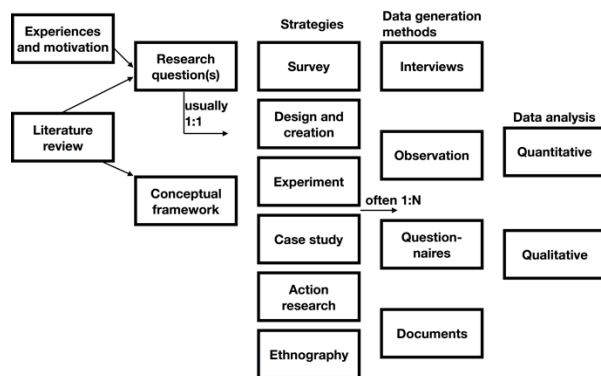


Figure 2: Model of the research process (Oates, 2005)

Oates (2005) also notes the process and components for doing research (Figure 2): “personal experiences and motivation, literature review, research question, conceptual framework, strategies, data generation methods and quantitative and/or qualitative data analysis” (Oates, 2005, p. 33).

As the many reasons to why people do research, so can the age distribution of PhD-students vary. Figure 3 presents the age distribution of PhD-students in a 2016 survey (The Swedish Higher Education Authority, 2016). Why is age important? It is because although the survey does not cover it, some doctoral students may have had a professional career, or are still practitioners in their professional field with many years of experience: these are also known as ‘pracademics’ (Dickinson, Fowler & Griffiths, 2020).

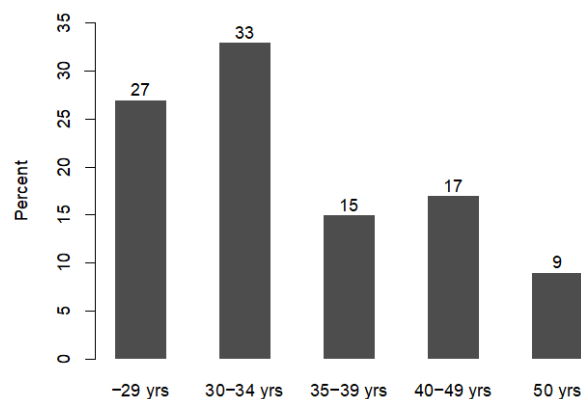


Figure 3: Age distribution (%) of PhD-students.

The author of this article falls within the 40-49 years age distribution. His reasons about *why* doing research on offensive cyberspace operations is based on his personal experiences during his professional career serving in the Swedish Armed Forces and the United Nations (UN), holding numerous intelligence and security positions including Director Intelligence in cyber-related issues and as Head of the UN's Intelligence Cell in a mission area for several years. In addition, he is a Veteran, with more than five years of duty in conflict and post-conflict areas including tours to the Balkans and in Central Asia. He was decorated with dual Non-Article 5 NATO Medals – and the Swedish Armed Forces International Service Medal of Reward for his tours to the Balkans and Central Asia.

These positions were held as a result of the professional training and education and achievements of each held role leading to measurable results. Having done a career with more than 20 years in the security and intelligence and interacting with numerous people ranging from those who had lost everything in a conflict to Ambassadors and Generals from various States and cultures, provide additional experiences which can never be gained by reading a book, or by taking a course. The book and the course may be a prerequisite to get a minimum threshold of understanding of the task at hand, but the experiences achieved after the engagement are based on that particular reality that existed at that particular time and space in that particular part of the world. Regardless of position held during these years, the problems have been complex. These complex problems included to provide a safe and secure environment (SASE) so people could build their lives, so economies could begin to grow, and governments could be built to govern, while in other areas to end conflict. As a security and intelligence professional, the purpose was to understand these complex problems, and based on this understanding, knowing the political landscape, identify the threats that were seeking to harm staff, assets and operations; identify the vulnerabilities that one self's organisation could have; and identify the associated risks. This was a continual collection of information, continual processing and analysis, and producing intelligence products in different formats. The observant reader notes that these steps are in a cyclical way, just like the intelligence cycle, and the problem, to collect information about the problem from different sources, to process, analyse and produce an intelligence product, reminds of the same steps as in Oates' (2005) model of the research process. Another model which is even more similar to the intelligence cycle in Figure 1 and widely used in research is Yin's (2018) case study research process (Figure 4).

3. The Research Project

The personal experiences and motivation for undertaking a doctoral project in offensive cyberspace operations is about understanding what offensive operations entail based on the author's 20+ years of experiences in a national and international setting as a security and intelligence professional, who's intelligence products are very similar to research products. The purpose with the project is to generate understanding to policy-makers, military leaders, academics and the general population, so Sweden can use offensive cyberspace operations for deterrence and defence: to defend Swedish national interests.

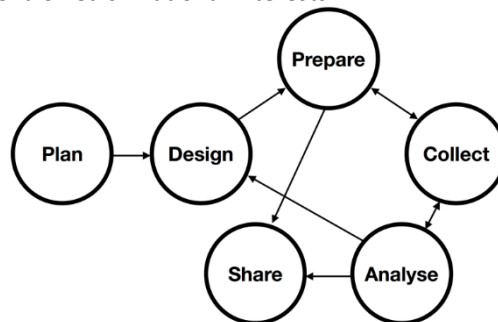


Figure 4: Case study research process (Yin, 2018)

The doctoral project consists of two components: an education part and a training part. The education part is about understanding the different research philosophies, and how these philosophies affect the approach to theory development, methodological choice, research strategies, potential time horizons, and techniques and procedures for data generation and data analysis. Saunders (2016) has dubbed this as the "research onion." The research project is based on three theories: deterrence theory, cyberspace operations theory and intelligence theory. From a philosophical point of view, a researcher can be standing on the research philosophy of positivism, interpretivism or pragmatism (Saunders, 2016). This affects the approach to theory development: deduction, abduction or induction, which in turn affects the methodological choice: mono quantitative or qualitative methods, multi-methods or mixed methods (Saunders, 2016). Finally, the choice of research

strategies (e.g., survey, action research, experiment, case study) affect the choice of data collection/data generation and analysis methods or techniques (Saunders, 2016).

Based on the personal experiences, motivation and philosophical stance (interpretivism), with related strategies (case study), data generation (e.g., interviews, observations) and analysis methods (e.g., thematic analysis), the training part of the doctoral project is about the planning and execution of research activities that lead to research products. The research products are double-blind peer reviewed, and after taking into account the reviewer's comments, the results are published in an acceptable venue. The output of this research project, using "The General Method of Theory-Building Research in Applied Disciplines" (Lynham, 2002), has been ten research articles published in conference proceedings and in one journal (Bengtsson & Huskaj, 2021; Granåsen, Huskaj & Varga, 2019; Huskaj & Moradian, 2018; Huskaj, 2019; Huskaj & Wilson, 2020a; Huskaj & Wilson, 2020b; Huskaj, Wilson & Iftimie, 2020; Huskaj & Wilson, 2021a; Iftimie & Huskaj, 2020; Kävrestad & Huskaj, 2021), one report (Huskaj & Sallinen, 2021d), and three unpublished works (Huskaj, 2018; Huskaj, 2021b; Huskaj 2021c). This article is not included in that count. Finally, an academic course (Cyber Operations in Antagonistic Environments) on advanced level at the Swedish Defence University is heavily based on the results of this research. It is *very likely* that these research products have been significantly influenced by this author's background, experiences, and expertise to identify a problem, plan, prepare, collect, analyse, produce and disseminate intelligence products: Figures 1, 2 and 4 show the close, if not equal, resemblance between intelligence operations and related work with scientific research.

4. The Method and Linguistic Analysis

The method to quantify 'substantial' is depicted in Figure 4. The limitations of this method include: if the assumption is different (e.g., the main author does less than 50%); qualitatively, if no principles, recommendations or guidelines exist in a particular field, it is not possible to apply linguistic analysis (Goldkuhl, 2002).

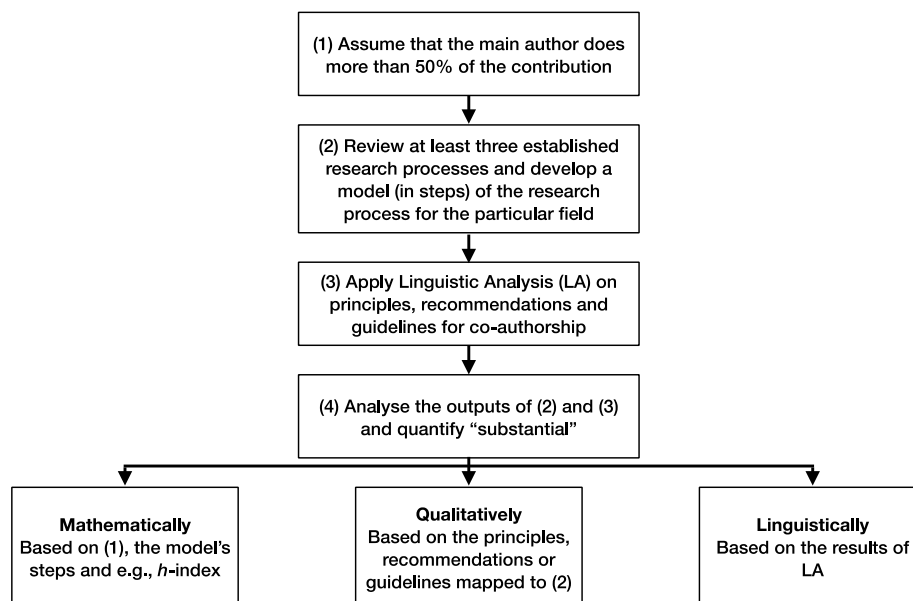


Figure 4: The method to quantify substantial

The doctoral project is done within a research discipline and supported by several supervisors. The supervisors supervise, a verb which means "to look over" and is "from Medieval Latin *supervisus*, past participle of *supervidere* 'oversee, inspect'" (Etymology Dictionary [herein ED], 2021b). In other words, the supervisors look over, oversee, and inspect the doctoral student's research activities that lead to research products published in an acceptable venue. However, to look over, oversee, and inspect, implies that certain formal requirements must be met. These formal requirements are, for the main supervisor, "at least decent competence is required" (HiS, 2017, p.2); for the primary and assistant supervisor, "at least a doctoral degree" (HiS, 2017, p. 2). These formal requirements are always met. Furthermore, while supervising (to look over, oversee, and inspect) the supervisor(s) have an important role to review (inspect) the doctoral student's work:

"Normally, research students and supervisors collaborate in the production of scientific work. It is therefore common for both research students and supervisors to be stated as co-authors of these works,

provided that the supervisor has participated in the production of the publication and / or the intellectual work on which the publication is based” (HiS, 2017, p. 5).

The above states that “research students and supervisors collaborate in the production of scientific work” and that this is “normally” done. *Collaborate*, a verb from 1845 means “to work with another or others” (ED, 2021c). *Production*, a noun, stems from *produccioun* meaning “a coming into being”, a “noun of action” (ED, 2021d). The verb “produce” is from the early 15th-century *producen*, to “develop, proceed, extend, lengthen out” (ED, 2021e). Finally, “normally” an adverb from the 1590s, means “regularly, according to general custom” (ED, 2021f).

In other words, according to general custom (=normally), those who look over, oversee, and inspect (=supervisors), work with another [e.g., the research student] or others, and take actions to put something into being (=production), or develop (an action, =produce) (e.g., the research product).

The next sentence states that the research student and those who look over, oversee, and inspect are “stated as co-authors of these works” (=research products). The word “author”, a noun, is from the mid 14th-century and stems from *auctor*, *autour*, *autor*, meaning “father, creator, one who brings about, one who makes or creates” (ED, 2021g). The Latin *auctour* means “promoter, producer, father, progenitor; builder, founder; trustworthy writer, authority; historian; performer, doer; responsible person, teacher” (ED, 2021g). The observant reader notes “producer”, a noun, from the 1510s which means “one who or that which produces” (ED, 2021h), an agent noun from the verb “produce”.

In other words, the research student and those who look, oversee and inspect are stated as “fathers, creators, [those] who bring about” the research product. The “one(s) who or that which produces” (=producer) take actions to develop (an action, =produce) a research product.

In conclusion, according to general custom (=normally), those who look over, oversee, and inspect (=supervisors), work with another [e.g., the research student] or others, and take actions to put something into being (=production), or develop (an action, =produce) (e.g., the research product), and the research student and those who look, oversee and inspect are stated as “fathers, creators, [those] who bring about” (=authors) the research product. The “one(s) who or that which produces” (=producer) take actions to develop (an action, =produce) a research product.

Now, consider “provided that the supervisor has participated in the production of the publication and / or the intellectual work on which the publication is based” (HiS, 2017, p. 5). The keyword here is “participated”, the past tense verb of “participate.” Participate, a verb, is from the 1530s and means “to partake, to share or share in” (ED, 2021i). The noun, participation, means “act or fact of sharing or partaking in common with another or others; act or state of receiving or having a part of something” (ED, 2021j).

In other words, those who look over, oversee and inspect (=supervisors) take actions (=verb) “to partake, to share or share in” (=participate) to develop a “written or printed work” (=publication) (ED, 2021k).

The observant reader notes that nothing is mentioned about how much actions are required by the research student and the supervisors to be considered as “creators” or “producers” (=authors) of the research product. In other words, “nothing is mentioned about the size of the contribution of the co-author in question” (NTNU, 2021). Because of this, co-authorship can lead to “frustration and conflicts” (NTNU, 2021). To combat frustration and conflicts, different principles, recommendations and guidelines have been produced. Offutt (2015) states that, as a general principle, “Everyone who makes substantial contributions to the results is a co-author on papers that present those results” (p. 73). NTNU (2021) states that, as a general principle, “decide who is a co-author and how the co-authors will contribute already at the beginning of the publication process”. Offutt (2015) seems to agree by stating that “it is very important that authorship must be discussed openly, objectively, and rationally. This conversation should happen before the paper is started” (p. 74). However, even Offutt (2015, p. 74) wonders: “*What is a substantial contribution?*” While Offutt (2015), based on his experience, provides a list (not comprehensive) of circumstances where authorship is warranted, the Vancouver Recommendations, drawn by the International Committee of Medical Journal Editors (ICMJE) (The Norwegian National Research Ethics Committees, 2020) provides the recommendation “that authorship be based on the following 4 criteria:

1. Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
2. Drafting the work or revising it critically for important intellectual content; AND
3. Final approval of the version to be published; AND
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.” (ICMJE, p.2).

This recommendation is also provided by several Swedish universities (e.g., Karolinska Institutet, Mid Sweden University, Umeå University; University of Skövde): “All those designated as authors should meet all four criteria for authorship, and all who meet the four criteria should be identified as authors” (ICMJE, 2019, p. 2). The keyword, as Offutt (2015) and ICMJE (2019) state is “substantial.” Substantial is an adjective and stems from the mid 14th-century, meaning “ample, sizeable” (ED, 2021l). Ample, an adjective, stems from the 12th-century “from Old French *ample* ‘large, wide, vast, great’” (ED, 2021m). In the 15th-century, it meant “great, abundant.” Sizeable, an adjective from the 1620s means “of relative good, suitable, or desirable size, usually somewhat large” (ED, 2021n).

In other words, *substantial*, means large, wide, vast, great (=ample) or of relative good, suitable, or desirable size, usually somewhat large (=sizeable). The Swedish equivalent “betydande” means “som spelar en stor roll, som utövar stort inflytande, framstående, vidgående” (SAOB, 2021), and is comparable to “ansenlig, avsevärd” (Svensk ordbok, 2021). The adjective substantial is about the size of something relative to something else, and the size is desirable, an adjective from the verb desire, “from the Latin *desidare* ‘long for, wish for; demand, expect’” (ED, 2021o). Thus, quantifying the size and deeming it desirable means that the entity that is longing for, wishing for, or demanding or expecting, has to express and describe it in qualitative and/or quantitative terms.

Applying the above in an academic research project with related activities where the output is a research product leads to the following reasoning. The output of a research project is a research product. The research product is commonly organised around abstract, introduction, literature review, methods, results and discussion. Conclusions may be omitted but are welcome if they bring further clarity to the topic. However, for a research product to come to life, substantial effort is required by the engaged researcher. Saunders (2016) “research onion” presents the steps in a research project: six in total. Goldkuhl (2020) presents the research activities as “formulating research questions, choose appropriate research strategies and methods for collecting / generating data, coding/analysing the data, concept formation (i.e., defining concepts), and the formulation of theoretical statements.” Goldkuhl (2020) presents six steps. Oates (2005) presents seven steps. However, while the steps may differ, they are not in a one-to-one-relation. Oates (2005) mentions “personal experiences, motivation and literature review” while the other two do not mention these. Both Goldkuhl (2020) and Saunders (2016) discuss the philosophical aspects of research, while this is not apparent in Oates’ (2005) research model. The remaining steps overlap.

The researchers have another step in common: the absence of the writing process. None of them mention the writing process in their models: the writing that is required to produce the research product that presents the results of the research project to a greater audience. However, Saunders (2016) has a whole chapter on “writing and presenting your project report” (pp. 628 - 673). Finally, none of the authors mention the review process a research product undergoes before publication.

Therefore, a more precis model of a research project is:

Personal experiences and motivation, understand the different research philosophies, and how these research philosophies affect the approach to theory development, methodological choice, research strategies, potential time horizons, and techniques and procedures for data generation and data analysis. Next, identify a problem and/or research idea, review the literature, conceptualise, generate a research question, generate a research aim and/or objectives, choose a relevant research design (which is “the general plan of how you will go about answering your research question(s)” (Saunders, 2016, p. 163)). The research design can be quantitative, qualitative or mixed methods (Saunders, 2016). A qualitative research design “is often associated with an interpretive philosophy” (Saunders, 2016, p. 168), has an abductive or inductive approach, and is linked to a qualitative data collection/generation technique/method such as semi-structured interviews (Saunders, 2016). The analysis is done using relevant data analysis methods, and the research strategy can be case study. After the

data analysis is done, it is then time to produce the first draft according to the steps as depicted in Table 1 (Borja, 2021). Note that although a qualitative research design was used to exemplify, all academic research can apply.

Table 1: The steps to produce the first draft of a research product

Step	Action	Step	Action
1	Prepare the figures and tables	7	Write the Abstract
2	Write the Methods	8	Compose a concise and descriptive Title
3	Write up the Results	9	Select Keywords for indexing
4	Write the Discussion. Finalize the Results and Discussion before writing the introduction. This is because, if the discussion is insufficient, how can you objectively demonstrate the scientific significance of your work in the introduction?	10	Write the Acknowledgements
5	Write a clear Conclusion	11	Write up the References
6	Write a compelling introduction		

Then, write a cover letter, upload the manuscript and send it to the Journal and respond to eventual comments sent by the editor and/or anonymous reviewers responsible for the double-blind peer review.

In summary, a model of a research project is as follows. Based on personal experiences and motivation, understand the different research philosophies, and how these research philosophies affect the approach to theory development, methodological choice, research strategies, potential time horizons, and techniques and procedures for data generation and data analysis. Table 2 presents a model of a research project in offensive cyberspace operations consisting of 27 steps.

Table 2: A model of a research project in offensive cyberspace operations

Step	Action	Step	Action
1	Identify a problem and/or research idea	15	Write a clear Conclusion
2	Review the literature	16	Write a compelling introduction
3	Conceptualise	17	Write the Abstract
4	Generate a research Question	18	Compose a concise and descriptive Title
5	Generate a research Aim and/or Objectives	19	Select Keywords for indexing
6	Choose a relevant research design (e.g., qualitative). A qualitative research design implies an interpretative philosophy with an abductive or inductive approach	20	Write the Acknowledgements
7	Choose a research strategy (the plan to achieve the goal). A qualitative research strategy can be Case Study	21	Write up the References
8	Choose one/several data collection/generation technique(s)/methods (e.g., semi-structured interviews, self-reporting)	22	Write a cover letter
9	Collect the information	23	Upload the manuscript
10	Analyse the collected data using qualitative procedures	24	Respond to the journal's decision, which may include:
11	Prepare the figures and tables	25	Respond to the editor
12	Write the Methods	26	Review the comments from the double-blind peer reviewers, and edit the manuscript where needed
13	Write up the Results	27	Re-send the manuscript
14	Write the Discussion. Finalise the Results and Discussion before writing the introduction. This is because, if the discussion is insufficient, how can you objectively demonstrate the scientific significance of your work in the introduction?		

What does this have to do with a substantial contribution? Recall that the adjective substantial is about the size of something relative to something else, and the size is desirable, an adjective from the verb desire, which is "from the Latin *desidare* 'long for, wish for; demand, expect' (ED, 2021o). Thus, quantifying the size and deeming it desirable means that the entity that is longing for, wishing for, or demanding or expecting, has to express and describe it in qualitative and/or quantitative terms. Thus, the total size of a research project in offensive

cyberspace operations consists of 27 steps. While some steps in the model require less effort, others do not. Data collection may go quickly, but the analysis may take time (e.g., manual instead of computational). Furthermore, while writing the first draft of a research product aimed for a journal may take 12 weeks (Belcher, 2009), the response to the journal's decision may take an equal amount of time depending on the outcome: e.g., desk reject or major revisions.

5. Quantifying Substantial

To quantify substantial mathematically without taking into account any temporal aspects could be to contribute by conducting 13 of the 27 steps. It is assumed that a substantial contribution does not equal to 50% of the work. To quantify substantial qualitatively could be that the first author conducts step 1-23, and the co-author 24-27. This approach assumes the co-author has significant experience in the academic field. To make this tangible, the *h*-index, which compares "the productivity and impact of published work of scholars" (Scopus, 2021), can be applied. Based on the author's experience, a co-author with an *h*-index of 38, *h*-index of 16, or *h*-index of 11 is expected to provide in less steps than a co-author with a lower *h*-index. A co-author with a lower *h*-index is expected to contribute more, but never 50% of the work.

Another way of quantifying substantial is by taking into account the Vancouver Recommendations as mentioned earlier. There are four criteria, and these four are bound by an AND boolean-operator. Therefore, the first criteria are about conception of the work (steps 1 - 5); or design of the work (steps 6 -8); or the acquisition (step 9); analysis (step 10); or interpretation of data for the work (step 11); AND Drafting the work (steps 12 - 16) or revising it critically for important intellectual content (steps 12 - 17); AND Final approval of the version to be published (review steps 12 - 21); AND Agreement of accountability.

The observant reader may note that the Vancouver Recommendations only take into account the first draft of the paper and does not cover steps 22 - 27. Finally, a mixed approach of the above scenarios can also be taken.

Therefore, the answer to the question "*What is a substantial contribution to a research project in cyberspace operations that merits co-authorship*" is: scientific actions taken by a co-author to ensure that a contribution to a research project is of desirable size relative to the total size of the research project which lives up to the wishes, demands or expectations of the first author where the output of the scientific actions is a research product. The related steps: cover letter, upload of the manuscript and responses to the journal's decision should preferably be discussed before the research project begins.

6. Conclusions and Why is this Important?

A PhD-project spans over several years and the psychological and emotional aspects coupled with e.g., busy supervisors, different views on co-authorship, and being inexperienced compared to the PhD-student, are all elements that can lead to frustration and conflict. The answer to the research question (as noted above) means a co-author in a research project in offensive cyberspace operations has to do some actual work where the output is a research product. Because PhD-students have different backgrounds (traditional, professional, pracademic), it is important for all of the involved parties in a PhD-project to understand that the needs, education and training may differ among the three mentioned categories of PhD-students. Senior academics should not be surprised if a professional PhD-student disagrees on co-authorship because the wishes, demands or expectations of the first author are not met. This is especially true if the supervisors take time to provide feedback, are not invested in the PhD-project (=being away for four years and wanting to invest 5% of one's time is not enough), not providing enough of support, being too busy (=2 years post-doc studies and an additional 1.5 years on a MSc-education), having different perspectives on co-authorship ("I had to write everything when I was a PhD-student and my professor just put his/her name on as co-author"), and being inexperienced in comparison to the PhD-student, are all elements that have led to no research products co-authored with the supervisors. To just review a draft manuscript and saying "this is good, now you can put my name on it" is not a substantial contribution and will not suffice for co-authorship.

Acknowledgements

I'd like to thank my former junior supervisors (in alphabetical order) Marcus Nohlberg and Johan Sigholm for inspiring me to write this article.

References

- Bengtsson, J. & Huskaj, G. (2021). The Manifestation of Chinese Strategies Into Offensive Cyberspace Operations Targeting Sweden. In: 20th European Conference on Cyber Warfare and Security / [ed] Dr. Thaddeus Eze, Dr. Lee Speakman and Dr. Cyril Onwubiko, Chester, 2021, p. 35-43.
- Belcher, W.L. (2009)., Writing your journal article in 12 weeks: A Guide to Academic Publishing Success. SAGE Publications.
- Borja, A. (2021). 11 steps to structuring a science paper editors will take seriously. Retrieved from <https://www.elsevier.com/connect/11-steps-to-structuring-a-science-paper-editors-will-take-seriously>.
- Dickinson, J. Fowler, A. & Griffiths, T.L. (2020). Pracademic expertise and its value to higher education. Retrieved from <https://www.advance-he.ac.uk/news-and-views/pracademic-expertise-and-its-value-higher-education>.
- Etymology Dictionary. (2021a). Motivation (n.). Retrieved from <https://www.etymonline.com/search?q=motivation>.
- ED (2021b). Supervise (v.). Retrieved from <https://www.etymonline.com/search?q=supervise>.
- ED. (2021c). Collaborate (v.). Retrieved from <https://www.etymonline.com/search?q=collaborate>.
- ED. (2021d). Production (n.). Retrieved from <https://www.etymonline.com/search?q=production>.
- ED. (2021e). Produce (v.). Retrieved from https://www.etymonline.com/word/produce?ref=etymonline_crossreference#etymonline_v_2617.
- ED. (2021f). Normally (adv.). Retrieved from https://www.etymonline.com/word/normally#etymonline_v_35603.
- ED. (2021g). Author (n.). Retrieved from https://www.etymonline.com/word/author#etymonline_v_18965.
- ED. (2021h). Producer (n.). Retrieved from https://www.etymonline.com/word/producer#etymonline_v_30401.
- ED. (2021i). Participate (v.). Retrieved from <https://www.etymonline.com/search?q=participate>.
- ED. (2021j). Participation (n.) Retrieved from https://www.etymonline.com/word/participation?ref=etymonline_crossreference.
- ED. (2021k). Publication (n.) Retrieved from <https://www.etymonline.com/search?q=publication>.
- ED. (2021l). Substantial (adj.). Retrieved from https://www.etymonline.com/word/substantial#etymonline_v_22270.
- ED. (2021m). Ample (adj.) Retrieved from https://www.etymonline.com/word/ample#etymonline_v_11020.
- ED. (2021n). Sizeable (adj.). Retrieved from https://www.etymonline.com/word/sizeable#etymonline_v_38192.
- ED. (2021o). Desire (v.). Retrieved from https://www.etymonline.com/word/desire?ref=etymonline_crossreference#etymonline_v_5657.
- Goldkuhl, G. (2002). Anchoring scientific abstractions – ontological and linguistic determination following socio-instrumental pragmatism. European Conference on Research Methods in Business and Management (ECRM 2002), Reading, 29-30 April 2002.
- Goldkuhl, G. (2020). Kvalitativ forskning: frågor, data, teori. Presentation.
- Granåsen, M., Huskaj, G., & Varga, S. (2019). Data Collection and Research in CDXs - Command and Control, Cyber Situational Awareness and Intelligence Perspectives on Cyber Defense. In 24th International Command and Control Research & Technology Symposium (p. 20). Laurel.
- HiS. (2017). Riktlinjer. Retrieved from <https://www.his.se/globalassets/dokument/styrdokument/riktlinjer-handledning-studieuppfoljning-utbildning-forskarniva.pdf>.
- HiS. (2019). Riktlinjer för vetenskapligt författarskap. Retrieved from <https://www.his.se/globalassets/tillfalliga-dokument/styrdokument/forskning-och-utbildning-pa-forskarniva/riktlinjer-for-vetenskapligt-forfattarskap.pdf>.
- Huskaj, G. (2018). A computational literature review of cyberspace operations. Unpublished work.
- Huskaj, G., & Moradian, E. (2018). Cyber Deterrence: An Illustration of Implementation. In J. Hurley & J. Chen (Eds.), 13th International Conference on Cyber Warfare and Security (pp. 304–311). Washington D.C.: Academic Conferences and Publishing International Ltd.
- Huskaj, G. (2019). The Current State of Research in Offensive Cyberspace Operations. In T. Cruz & P. Simoes (Eds.), 18th European Conference on Cyber Warfare and Security (pp. 660–667). Academic Conferences and Publishing International Ltd.
- Huskaj, G., & Wilson, R. L. (2020a). An Anticipatory Ethical Analysis of Offensive Cyberspace Operations. In P. B. K. Payne & P. H. Wu (Eds.), ICCWS 2020 15th International Conference on Cyber Warfare and Security (pp. 512–520). Norfolk: Academic Conferences and Publishing International Limited.
- Huskaj, G. & Wilson, R.L. (2020b). Anticipatory Ethics for Vulnerability Disclosure. In: Proceedings of the 15th International Conference on Cyber Warfare and Security / [ed] Payne, Brian K.; Wu, Hongyi, Reading, UK: Academic Conferences and Publishing International Limited, 2020, pp. 254-261.
- Huskaj, G., Iftimie, I. A., & Wilson, R. L. (2020). Designing attack infrastructure for offensive cyberspace operations. In European Conference on Information Warfare and Security, ECCWS (Vol. 2020-June, pp. 473–482).
- Huskaj, G., & Wilson, R. L. (2021a). Offensive Cyberspace Operations and Zero-days: Anticipatory Ethics and Policy Implications for Vulnerability Disclosure. *Journal of Information Warfare*, 20(1), 96–109.
- Huskaj, G. (2021b). How do Smart States Organize Cyberspace Capabilities? The Swedish Case. Unpublished work.
- Huskaj, G. (2021c). Testing a Framework for Offensive Cyberspace Operations. Unpublished work.
- Huskaj, G. & Sallinen, M. (2021d). Staters outtalade normer i cyberrymden. Retrieved from <http://fhs.diva-portal.org/smash/get/diva2:1576646/FULLTEXT02.pdf>.
- ICMJE. (2019). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Retrieved from <http://www.icmje.org/icmje-recommendations.pdf>.

- Iftimie, I.A. & Huskaj, G. (2020). Strengthening the cybersecurity of smart grids: The role of artificial intelligence in resiliency of substation intelligent electronic devices. In: Proceedings of the 19th European Conference on Cyber Warfare and Security: a virtual conference hosted by University of Chester UK 25-26 June 202 / [ed] Thaddeus Eze, Lee Speakman, Cyril Onwubiko, Reading: Academic Conferences and Publishing International Limited, 2020, pp. 143-150.
- Joints Chiefs of Staff. (2013). Joint Publication 2-0: Joint Intelligence. Retrieved from https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf.
- Kävrestad, J. & Huskaj, G. (2021). How the Civilian Sector in Sweden Perceive Threats from Offensive Cyberspace Operations. In: 20th European Conference on Cyber Warfare and Security / [ed] Dr. Thaddeus Eze, Dr. Lee Speakman and Dr. Cyril Onwubiko, Chester: Academic Conferences and Publishing International Limited, 2021, pp. 499-506.
- Morrison-Saunders, A., Moore, S.A., Hughes, M. and Newsome, D. (2010). *Coming to terms with research practice - Riding the emotional rollercoaster of doctoral research studies*. In: Thomson, P. and Walker, M., (eds.) The Routledge doctoral student's companion: getting to grips with research in education and the social sciences. Routledge, New York.
- Lynham, S.A. (2002). The General Method of Theory-Building Research in Applied Disciplines. *Advances in Developing Human Resources* 2002; 4; 221 DOI: 10.1177/1523422302043002.
- NTNU. (2021). Co-authorship. Retrieved from <https://innsida.ntnu.no/wiki/-/wiki/English/Co-authorship>.
- Nelson, M.L. & Friedlander, M.L. (2001). A Close Look at Conflictual Supervisory Relationships: The Trainee's Perspective. *Journal of Counseling Psychology*. 2001, Vol. 48, No. 4, pp. 384-395.
- Nygaard, L.P. (2017). Publishing and perishing: an academic literacies framework for investigating research productivity. *Studies in Higher Education*, 42:3, pp. 519-532, DOI: 10.1080/03075079.2015.1058351.
- Oates, B. J. (2005). *Researching Information Systems and Computing* (First). Sage Publications, Inc.
- Offutt, J. (2015). Editorial: Who Is An Author?. *Softw. Test. Verif. Reliab.* 2015; 25:73–76. DOI: 10.1002/stvr.1571.
- SAOB. (2021). Betydande. Retrieved from https://svenska.se/saob/?id=B_2123-0004.wdK4&pz=3.
- Saunders, M. N. K., Lewis, P., & Thornhill, A. (2016). *Research Methods for Business Students* (7th ed.). Harlow: Financial Times/Prentice Hall.
- Scopus. (2021). Scopus. Retrieved from <https://www.scopus.com>.
- Svensk ordbok. (2021). Betydande. Retrieved from <https://svenska.se/so/?sok=betydande&pz=4>.
- The Norwegian National Research Ethics Committees. (2020). The Vancouver Recommendations. Retrieved from <https://www.forskningsetikk.no/en/resources/the-research-ethics-library/legal-statutes-and-guidelines/the-vancouver-recommendations/>.
- The Swedish Higher Education Authority. (2016). Doktorandspegeln – en enkät om doktorandernas studiesituation. Retrieved from <https://www.uka.se/publikationer--beslut/publikationer--beslut/rapporter/rapporter/2016-12-08-brister-i-doktoranders-handledning.html>.
- Times Higher Education. (2017). Of monsters and mentors: PhD disasters, and how to avoid them. Retrieved from <https://www.timeshighereducation.com/features/monsters-and-mentors-phd-disasters-and-how-to-avoid-them>.
- Tschannen-Moran, M. & Nestor-Baker, N. (2004). "The Tacit Knowledge of Productive Scholars in Education." *Teachers College Record* 106 (7): 1484–511.

Human Errors: A Cybersecurity Concern and the Weakest Link to Small Businesses

Tabisa Ncubukezi

Information Technology Department, Faculty of Informatics and Design, Cape Peninsula University of Technology, Cape Town, South Africa

ncubukezit@cput.ac.za

Abstract: Cybersecurity is essential for all organizations, especially during this menacing Covid-19 global pandemic. The sudden transition of leaving the offices to work from home – the 'new normal' – has introduced information security-related risks associated with human factors. For example, both criminals and employees use the same platform for information exchange but with starkly different intentions. But both their actions compromise information and computer security. Criminals intentionally exploit systems to gain unauthorized access for their benefit, while employees make careless mistakes, leaving systems exposed and vulnerable. The present study examines human errors influenced by actions, attitudes, and behaviors that affect overall information security. Purposive sampling within the qualitative approach was used to select thirty (30) small business managers. Data was collected using a qualitative online survey as a Google Form. The study used thematic analysis. The results revealed that repeated human mistakes compromise information security principles and render employees the weakest link. The study explained the risks caused by employees due to ignorance or poor decision making, technical-related errors, and skills- and policy-based errors. Even though small businesses do not require a 'one-size-fits-all' security approach, recommendations to reduce human mistakes were made.

Keywords: Computer security, cybersecurity, cyber threats, data breaches, human errors, information security, skills-based errors, small businesses

1. Background of the study

The state of cybersecurity as an ongoing topic has gained popularity in academia. The literature presented the state of cybersecurity, common cyber threats, their impact on businesses, and technology-related issues (Akhtar, Sheorey & Bhattacharya, 2021). Studies expose increased cybercrimes that affect information security, economic and business growth (Tam, Rao & Hall, 2021). For example, during the IMB security week in 2020, Sizwe Dlamini stressed the financial impact caused by cyber incidents, estimated to be R40.2 million per breach. And the estimation of the study fell short of including all organizations in South Africa (SA). Many studies have explored security threats and attacks initiated by cybercriminals. As a result, most studies report that increased data breaches and loss of information among businesses are associated with cybercriminals. This present study's interest is on internal users rather than criminals. Schneier (2000) highlights that people have become the weakest link in the chain of handling information, despite that employees tend to believe that their actions on a system cannot yield to security breaches (Hadlington, 2017; Herath & Rao, 2009).

On the contrary, several studies reveal the ignorance of business employees who often become an entry for cybercrimes (Sasse, Brostoff & Weirich, 2001). Solvere (2021) believes that cyber-attacks can be generated inside or outside an organization, rendering employees the most significant threat. The main cyber-attacks are, in fact, from human actions (Annarelli, Nonino & Palombi, 2020). Despite numerous studies on the impact of human factors on cybersecurity, human factors remain at the core of many cybersecurity breaches. As in this study, humans are business employees who carry out business processes to achieve business goals. Some studies refer to employees as the weakest link because of their power in exposing organizations (Annarelli, Nonino & Palombi, 2020). The term links employee negligence in making decisions that weaken an organization's information security (Teufel et al., 2020).

Human mistakes can frequently result from a convenient use of shortcuts to make employees' lives easier while exposing businesses to threats. Such incidents qualify the human element as a channel for attacks and a significant source of data breaches (Kobis, 2021). This research is presented alongside the alarming increase in cyber incidents caused by human behaviors within the small business space. This paper responds to the study of Ncubukezi, Mwansa, and Rocaries (2021), explaining that human-generated mistakes should be acknowledged when planning and implementing best practices to promote good cyber hygiene. The current work examines human errors fuelled by attitudes, actions, and behaviors that have emerged recently as a serious concern and a door to increased data breaches. This paper further presents the common types of human errors, their impact, and vulnerability mitigation strategies to improve the overall security of information processing.

The rest of the paper is structured as follows: cybersecurity relating to human factors, followed by the method of inquiry, results, and discussions. The final section presents the recommendations, contributions, future study suggestions, and concluding remarks.

2. Cybersecurity

Cybersecurity has many dimensions – cyberspace, information security, human factors, and computer security – necessitates that organizations identify loopholes and protect themselves from various cybercriminals (Ncubukezi & Mwansa, 2021). This study focuses on human factors, which presently receive less attention than technological or policy aspects related to security (Ergen, Ünal & Saygili, 2021). During the process of information flow, the human element is the key role player. Human errors are accidental actions influenced by attitudes, knowledge, and behaviors that expose businesses to quantified cyber threats and attacks. Figure 1 illustrates how employees work within the information flow to carry out their duties. The figure shows employees, services, a firewall to filter incoming and outgoing traffic, Internet connection, and the perpetrator. When proper security measures and firewalls exist with appropriate applications, employees receive error-free services. Perpetrators are unable to find a way to penetrate a system.

However, criminals use any opportunity to gain unauthorized access when security measures are absent, resulting in significant damage. This loophole is due to cyberspace and its nature, which is convenient, open, and grants equal access to all users (Ncubukezi, Mwansa & Rocaries, 2020). So access to timely and open cyberspace allows for cyber-attacks through user actions depending on safety measures applied.

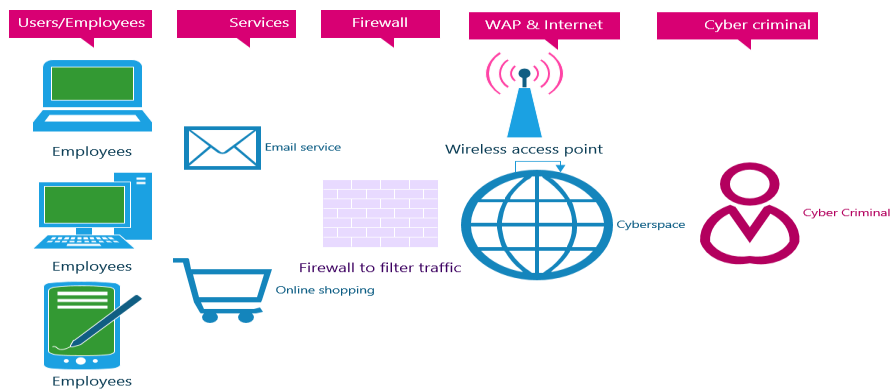


Figure 1: User processes during the information processing (Source: Own work, 2021)

2.1 Security risks posed by human errors

Human factors are regularly underestimated and overlooked (Hadlington, 2017) and are vital factors that affect a business's information security hygiene (Anwar et al., 2016). The challenge stems from the diverse range of human errors which ultimately grant unauthorized access to sensitive information and other business assets, resulting in significant data and security breaches. Employee mistakes pose a risk in companies. In fact, the rise and severity of security problems reported in recent years suggest that organizations are more vulnerable than ever (Sasse, Brostoff & Weirich, 2001). As illustrated in Figure 2, human errors are influenced by certain attitudes, behaviors, and actions that promote unsecured connections. These ignorant actions expose valuable, sensitive business information and resources to opportunistic criminals. Criminals then hijack secure sessions to violate privacy (Wallace et al., 2021). When cybercriminals take over, they compromise information security principles like data confidentiality, availability, and integrity. Confidentiality, a fundamental principle to promote protection against unauthorized disclosure of data or information, focuses on keeping information private. Data is only available to or can only be accessed by the correct recipient to carry out expected duties (Njoroge, 2020). The confidentiality principle includes people protecting others by restricting personal or sensitive information sharing unless explicit permission is granted (Alexei & Alexei, 2021).

The second principle focuses on the protection of data or information to promote its integrity. Business data should not be accessed for unauthorized modification, additions, or deletions. The integrity principle promotes trust and accuracy so that data remains the same (Angafor, Yevseyeva & He, 2020). Consequently, integrity encourages adherence to ethical tenets and fairness (Alexei & Alexei, 2021). Criminals typically take chances

when an employee acts negligently on a system, thereby opening a door for data leakage (Ergen, Ünal & Saygili, 2021).

The third and final principle protects and promotes system functionality by ensuring that data is always available for authorized users (Zimmermann & Renaud, 2019). Any ignorant action of the user could expose the system and information to attackers who compromise the principle, resulting in poor service and delayed decision making.

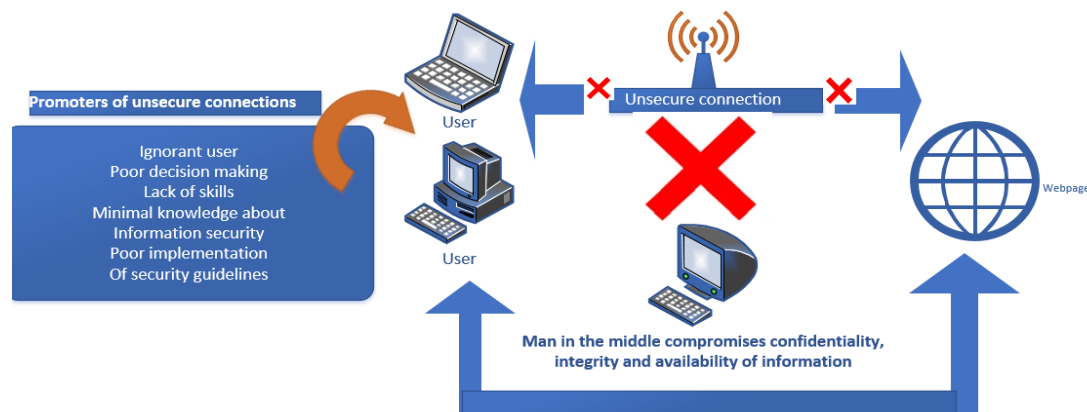


Figure 2: Illustration of the man in the middle attack (Source: Own work, 2021)

Given the above, human actions can undoubtedly affect the primary principles of information security, compromising business information and computer security. Information security principles are affected by threats associated with human factors such as social engineering, malware, phishing, worm, and spyware (Karaci, Akyüz & Bilgici, 2017). Alexei and Alexei (2021) insist that threats related to spoofing, access to unauthorized data lead to data theft and malicious programs. The implications of compromised information security as the result of human factors are presented below.

2.2 Common human mistakes

Cybersecurity risks relating to human mistakes affect various businesses because of the connection to standalone or networked computers. Moreover, Kobis (2021) believes that the human factor is the leading factor in infiltrating sensitive information. For standalone computers, employees may use memory sticks which a virus may infect. Or, for example, employees may follow a website's links or accidentally respond to unknown links that gather sensitive information. Increasingly, data breaches occur through the unauthorized disclosure of personal information (Richardson et al., 2020). Another example is when users curiously, recklessly, and ignorantly open fake emails containing malware attachments which automatically installs when opened.

Moreover, a user could install malware attached to standard applications. Often the infected installation package is available on a website to trap unknowledgeable users. In this case, the unaware user downloads and installs software from unverified sources (Kobis, 2021). Other users serve as a channel for criminals by the way they handle their passwords. Such behavior may result from a poor ability to remember accepted characters for password criteria, understaffing, and employee overload with work demands. At times, unacceptable user behavior is exacerbated by a lack of support or the absence of relevant training. This mistake is a gap that influences poor decision-making (Sasse, Brostoff & Weirich, 2001). Some user attitudes affect common mistakes; for example, when a user insists, "It won't happen to me" (Richardson et al., 2020).

3. Research method

This qualitative interpretive study used purposive sampling to select the research participants most likely to share valuable and appropriate information about the study. The sample comprises business managers representing the population of the study. The chosen sampling method aids the researcher in gaining insightful information based on participant knowledge and experience regarding human errors in small businesses. Furthermore, the process allows the researcher to detail the significant impact of the research findings.

Participants: The research participants had more than two years of experience and provided insight into the current study. A total of thirty (30) respondents were selected from the supply chain business sector, and each

participating business had a range of twenty (20) to eighty (80) employees. The research participants comprised a combination of males and females between 38 and 50 years of age.

Selection criteria: This study focused on small information and communication technology (ICT) businesses. The ICT business sector has been chosen based on its dependency on ICT resource usage, especially during the global Covid-19 pandemic. The criteria for selecting these businesses are twofold: 1) that companies have a range of 1 to 50 employees, and 2) that companies have a turnover of less than or equal to R10 000 000 a year.

Recruitment of participants: Research participants were invited to a virtual Zoom workshop for small businesses. The researcher followed up with an email, a consent form, and a link to the online survey. All participants receiving the email took part in the study and completed the questionnaire, giving a 100% response rate. All respondents assured their details and the data collected would remain private, confidential, and anonymous.

Data collection method: The study used an online qualitative questionnaire to gather data from knowledgeable and experienced participants. The current Covid-19 pandemic has escalated the use of online surveys on Google Forms. The questionnaire contained closed-ended questions (such as participant position, number of years in the company, number of employees in the business). In addition, open-ended questions were asked to ascertain the participants' awareness and their businesses about various types of human errors, the impact of the errors, and their businesses' current vulnerability mitigation strategies.

Analysis of data: This research used thematic analysis to uncover relevant themes by analyzing patterns, differences, and similarities in the collected data. This flexible, interpretive data analysis method assists a researcher in investigating participants' points of view, searching for differences and similarities (Braun & Clarke, 2006). The analysis method involved data preparation by searching across data sets to identify data, analyze pieces of data, and interpret data in a meaningful manner for the presentation of data. The process allowed the researcher to categorize data using the visual color-coding method for the initial codes generation. The researcher then prepared and organized the collected data using the visual method to search for themes. Data was then reviewed and explored to validate the themes using a thematic map. The data were categorized and interpreted on a board to search for patterns and similarities to define the themes. A final report of the findings was generated based on the interpretation of the collected data. The following themes emerged and were cohesively presented:

- factors influencing human errors;
- impact of the human errors; and
- vulnerability mitigation strategies.

4. Results and discussions

The current study examined human errors influenced by employee actions, behaviors, and attitudes that negatively influence the state of information and computer security. Every effort performed by employees on a system results in either a positive or a negative impact. This study demonstrated the different experiences relating to human factors that threaten information resources in information management. This research pertains to small businesses in the supply chain industry.

Table 1 shows several questions asked to gather data relating to types of human behaviors in information security and their root causes. These questions are based on employee experiences that influence unauthorized access to a business system. Human actions on business systems are generally shaped and influenced by the nature of the business environment, the level of awareness, and the enforcement of existing policies. Employee ignorance and decision-making typically cause these actions: inadequate skills for interacting with the system, related technical errors, and policy-based errors. The table's four columns present questions based on types of human error, followed by employee actions related to human errors. The table also has 'yes' and 'no' columns. Those businesses that experienced employee activities that negatively impacted the system are presented on the 'yes' column, while those which have not experienced employee mistakes are in the "no" column. Both the 'yes' and 'no' columns show the percentage of participant responses.

Table 1: Questions asked that relate to the employee errors

Questions based on the human error type	Employee actions relating to the error type	Yes	No
1. Do you experience human errors relating to employee ignorance and poor decision-making?		81%	19%
2. What are the employee actions relating to employee ignorance and poor decision-making?	- Deleting important files without a clear understanding of their value - Sending emails or documents with sensitive data to incorrect recipients or public platforms - Making unnecessary changes in documents - Sending sensitive business data via unsecured platforms	58%	42%
3. Do you experience skill-related to human errors?		36%	64%
4. What are the employee actions relating to the application of skills?	- Lack of awareness due to employee age - Following and opening email links and attachments from unknown recipients - Lack of proper understanding about cybersecurity - Using unsecured connections when sending sensitive data	46%	54%
5. Do you experience technical-related errors?		51%	49%
6. What technical errors have you experienced?	- Misconfigured business assets to permit unauthorized access - Lack of backup systems for critical data - Ignoring software updates - Downloading unauthorized software applications - Performing unauthorized system changes	28%	72%
7. Do you experience policy-based human errors?		80%	20%
8. What are the causes of those policy-based human errors?	- Poor handling of passwords and sensitive data - Misuse of access of rights and privileges	73%	27%

4.1 Factors influencing human errors

As illustrated in **Table 1**, participants shared their different experiences, including their take on human errors and how employees make those errors. Businesses have become victims of human error, and human mistakes are one of the biggest challenges for small businesses (Kobis, 2021). Several respondents confirmed the notion of the "human as the weakest link." With the growing rate of common mistakes by the employees within the business space, Kobis (2021) suggests that the cybersecurity behaviors of employees should be scrutinized as they form a significant part of the business. Various negative actions and employee behaviors are detrimental to the business sector, especially in organizations' private information and financial aspects (Ergen, Ünal & Saygili, 2021).

The results revealed that some human mistakes are caused by employee ignorance and poor decision-making, which Turk (2013) associates with staff shortage, fatigue, or working speed. Every business has a technical side that should not be ignored, and employees need to be equipped to understand technical terms (Campean, 2019). The business sector should engage a vigilant management team to enforce policies and security measures for information processing, people, hardware, and software (Greavu-Serban & Serban, 2014). The absence of a dedicated cybersecurity team and management involvement poses a risk, increasing poor adherence to and enforcement of policies and guidelines.

4.2 Impact of the human errors

Table 2 presents the questions regarding the impact of human errors within the small business sector. The effect of human errors has increased significantly during Covid-19. The study posed the following questions to gain an understanding of the effect of human behaviors.

Table 2: Impact of the human errors

Questions based on the impact of human errors
1. How have human errors during the global Covid-19 pandemic affected your business?
2. What is the primary area of the business negatively impacted as a result of human errors?
3. What are the main influencers of and contributors to human errors?
4. What employee actions relate to the application of skills?

The sudden Covid-19 global pandemic introduced changes in all institutions and persons. Everyone relies on the Internet as the backbone for communications and other daily operations during this present season. With increasingly convenient use and dependency on the Internet, criminal innovation has spiked and attempted new strategies. Sometimes, human mistakes serve as a channel for attracting a diverse range of attacks and threats that negatively affect the system. When asked about *the impact of human errors within the business during the pandemic*, the results revealed various experiences. Fifty-two percent of the respondents indicated a loss of data; 49% admitted to a loss of client trust; 28% were affected by psychological risks of managers; and 46% indicated societal implications, which eventually marred the business reputation. Sixty-eight percent of the respondents further indicated that the consequences of human errors resulted in the loss of money and investors; 68% experienced business disruption, and 60% indicated that their business eventually verged on discontinuing. Participants believe that human errors pose a significant threat to businesses. The primary sources of human errors are employees' diverse attitudes, behaviors, and actions influenced by employee moods and feelings.

According to participants, various reasons influence employee actions, attitudes, and behaviors. For example, participants mentioned that human errors are caused by the following: downloading suspicious attachments; following unknown links; minimal data backup systems; employee ignorance; employee attitudes; poor decision making; lack of cybersecurity skills and awareness; failure to use strong passwords; unlimited user access privileges; and other technical-related errors. Ignorance and poor adherence to mitigation measures cost businesses money, investors, growth, and societal impact resulting in a poor reputation, business disruption, data loss, client loss, and eventually business discontinuity (Ncubukezi, Mwansa & Rocaries, 2020). Human errors cause significant damage to businesses and remain a primary security concern. Cyber-attacks result in substantial economic harm that ruins businesses (Teufel et al., 2020). One participant believes that *"risks are not for small businesses."* Another respondent suggests that *"taking security measures should not be the employee responsibility; instead, it should be the company's responsibility."* Results indicate that human attitude contributes to the actions of the system. Unsurprisingly, 50% of the respondents confirmed that *"applying safety measures is time-consuming."* In contrast, 37% of the participants did not regard human errors as contributing to business risks. Some participants admitted that they are a human *"who is bound to make mistakes."*

4.3 Vulnerability mitigation strategies

Table 3 lists the questions about mitigation strategies that small businesses are adopting to protect their assets. Various participants shared their security measures to guard against unauthorized access.

Table 3: Security measures used

Questions based on the security measures used
1. How is the state of your current mitigation strategies?
2. How are the systems continuously updated?
3. What are the current protection measures used?
4. What are the employee actions relating to the application of skills?

With all the apparent human-related challenges, participants were queried about the *state of the current mitigation strategies*. Most participants indicated that their businesses currently have mitigation strategies even though proper implementation is not guaranteed. As one participant explained, *"We currently do not have a dedicated person to enforce the creation of strong passwords."* When asked about the continuous updates of the system and software, 62% of the respondents admitted to not running automatic updates because they feel the process consumes time. One respondent explained that *"sometimes we do not run the system updates as they delay us on our activities."* When asked about *current protection measures*, the respondents identified measures such as antivirus software and passwords. Only 78% of the respondents claimed to back up their information, while 22% believed their data was *"protected and out of reach."* All participants agreed that their primary security measure is based on the use of username and password. The nature of cybersecurity involves

humans – both legitimate (users) and illegitimate (criminals) – fighting for access to information and other resources (Assante & Tobey, 2011). Torton, Reaiche, and Boyle (2018) assert that as it is challenging to manage and guard human behavior, mitigation strategies are essential for every business. They further suggest that users should consistently practice safe and secure behaviors on the system.

Ncubukezi, Mwansa, and Rocaries (2021) explain that good security hygiene results from consistent, effective mitigation strategies. The continuous protection of business assets reduces human mistakes which threaten a business (Kobis, 2021). Many business systems are vulnerable to a diverse range of attacks and require proactive security measures that guard against every form of threat. Then, it is essential to continuously improve information security awareness (Aldawood & Skinner, 2018) by introducing programs that address strategies to protect against data breaches and reduce the likelihood of human-related risk (Alshaikh & Adamson, 2021). Unskilled users are most likely to make skills-related and technological errors from their lack of awareness and use of jargon. The absence of centralized account management is a significant challenge requiring an immediate solution. This risk is associated with policy-based errors, which arise when no detailed guidelines steer the use of a management database. As a result, poor management of user accounts and passwords opens a door for attackers (Greavu-Serban & Serban, 2014). Highly skilled employees promote good cyber hygiene for business systems (Lee & Rid, 2014).

The diverse business experiences reveal minimal practices of strategies that promote effective cybersecurity. As deliberated in section 2 of the paper, cybersecurity should be implemented on all levels. It becomes essential for every business to implement effective protection measures to decrease exposure to threats. Cybersecurity should be enforced and regularly reviewed for its effectiveness on all levels.

4.4 Recommendations

This paper forms part of ongoing Ph.D. work. With previous papers also linking to the main research, this study explores the business sector's human aspect. This study adopted best practices that should be taken into account (see Kayumbe & Michael, 2021; Coventry et al., 2014): use and management of passwords; use of updated antivirus and firewalls; use of updated software; shutting down of a computer; use of secure connections and websites; avoidance of phishing scams; limiting personal information; continuously checking physical surroundings when online, and reporting cybercrimes. Even though the 'blanket approach' is not overtly practical or serves all businesses, this study suggests general measures promote healthy systems.

Employees: It is necessary to understand the importance of cybersecurity. There should be regular training and awareness programs about essential security topics (including real-world scenarios), encouraging regular discussions, creating open platforms for questions, using daily security reminders, and monitoring employee activities. These security practices will keep users updated and abreast of the latest trends of attacks, thereby diminishing the chances of privacy violations.

Management: Business leaders should make informed decisions to set clear policies, rules, guidelines, and procedures for enforcing cybersecurity compliance; for example, enforcing the generation of passwords with adequate characters or enforcing two-factor authentication to reduce the chances of unauthorized access on a business network. Two-factor authentication is an electronic technique that increases security on applications or websites by enabling multiple authentications per session. It is one of the authentication mechanisms which grants access only after multiple verifications.

Information: Businesses must back up their systems daily to recover business data if a system is compromised, control physical access, and improve the security of payment processing.

System and technical: Businesses must limit access rights and privileges, install and update methods to protect against network attacks, including firewalls and encryption methods, conduct ongoing vulnerability testing on networks, and implement tools to scan networks and applications to detect a breach automatically.

5. Concluding remarks

Security of information and computers is a challenging task for any organization. While most research focuses on cybersecurity's technological and policy aspects, fewer studies have investigated human factors. This study reveals the impact of human error influenced by employee behaviors, attitudes, and uncontrolled actions in cyberspace. These activities have the potential to weaken business systems. It cannot be denied that human error is the primary contributing factor of cyber risk in the small business sector. In these uncertain times, cybercriminals use every opportunity to gain access to business systems.

Consequently, all businesses must resort to solutions that promote the longevity of the company. This work reports the causes and types of human errors and small business employee human error impact. This work also presents the recommended strategies for reducing cyber risks relating to human errors. The outcome discovered a range of factors relating to human error that compromise the privacy and security of business systems.

As a contribution: This work shares insights and escalates awareness of unintentional actions of small business employees. The study further shares insights about employee ignorance, which leads to poor decision-making, deficiency of awareness, understanding, and skills, and the necessary implementation of information procedures and computer security guidelines.

In the future: The researcher should explore employees' intentional actions (inside attempts) that result in successful data breaches within the small business sector. Furthermore, a study should include other business sectors. It would be of good value for the researcher to embed artificial intelligence to reduce risks.

Acknowledgements

I thank the reviewers for the time spent generating the feedback. I acknowledge and thank National Research Foundation (NRF) Black Academics Advancement Programme (BAAP) for financing this work.

References

- Akhtar, S., Sheorey, P.A. and Bhattacharya, S. 2021. Cyber Security Solutions for Businesses in Financial Services: Challenges, Opportunities, and the Way Forward. *International Journal of Business Intelligence Research*, 12(1), 82-97.
- Aldawood, H. and Skinner, G. 2018. Educating and raising awareness on cyber security social engineering: A literature review. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 62-68.
- Alexei, A. and Alexei, A. 2021. Cyber Security Threat Analysis in Higher Education Institutions As A Result Of Distance Learning. *International Journal of Scientific & Technology Research*, 10(3), 129-133.
- Alshaikh, M. and Adamson, B. 2021. From awareness to influence: toward a model for improving employees' security behavior. *Personal and Ubiquitous Computing*, 1-13.
- Angafor, G.N., Yevseyeva, I. and He, Y. 2020. Game-based learning: A review of tabletop exercises for cybersecurity incident response training. *Security and privacy*, 3(6), e126.
- Annarelli, A., Nonino, F. and Palombi, G. 2020. Understanding the management of cyber-resilient systems. *Computers & Industrial Engineering*, 149, 106829.
- Anwar, M., He, W., Ash, I., Yuan, X., Li, L., and Xu, L. 2017. Gender difference and employees' cybersecurity behaviors. *Computers in Human Behavior*, 69, 437-443.
- Assante, M.J. and Tobey, D.H. 2011. Enhancing the cybersecurity workforce. *IT Professional*, 13(1), 12-15.
- Braun, V. and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Campean, S. 2019. The Human Factor at the Center of a Cyber Security Culture. *International Journal of Information Security and Cybercrime (IJISC)*, 8(1), 51-58.
- Coventry, L., Briggs, P., Blythe, J. and Tran, M. 2014. Using behavioral insights to improve the public's use of cyber security best practices. *Gov. UK Report*.
- Dlamini, S. 2020. *Data breaches cost SA companies R40.2 million on average in 2020*. Available from <https://www.iol.co.za/business-report/companies/data-breach-costs-sa-companies-r402-million-average-in-2020-6649ae0a-b803-482c-978f-b395517c7fa7> [accessed on 12 September 2021].
- Ergen, A., Ünal, A.N. and Saygili, M.S. 2021. Is It Possible to Change the Cyber Security Behaviours of Employees? Barriers and Promoters. *Academic Journal of Interdisciplinary Studies*, 10(4), 210.
- Greavu-Serban, V. and Serban, O. (2014). Social engineering is a general approach. *Informatica Economica*, 18(2), 5.
- Hadlington, L. 2017. Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviors. *Heliyon*, 3(7), e00346.
- Herath, T. and Rao, H.R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems*, 47(2), 154-165.
- Karaci, A., Akyüz, H.I. and Bilgici, G. 2017. Üniversite öğrencilerinin siber güvenlik davranışlarının incelenmesi. *Kastamonu Eğitim Dergisi*, 25(6), 2079-2094.
- Kayumbe, A. and Michael, L. 2021. Cyber threats: Can Small Businesses in Tanzania outsmart Cybercriminals? *International Research Journal of Advanced Engineering and Science*, 6(1), 141-144.
- Kobis, P. 2021. Human factor aspects in information security management in the traditional IT and cloud computing models. *Operations Research and Decisions*, 31(1), 61-76.
- Lee, R.M. and Rid, T. 2014. OMG Cyber! Thirteen Reasons Why Hype Makes for Bad Policy. *The RUSI Journal*, 159(5), 4-12.
- Ncubukezi, T., Mwansa, L., and Rocaries, F. 2020. A Review of the Current Cyber Hygiene in Small and Medium-sized Businesses. *International Conference for Internet Technology and Secured Transactions (ICITST)*, 15, 283-288. IEEE.
- Ncubukezi, T., Mwansa, L., and Rocaries, F. 2021. Analysis and Impact of the Cybercrimes in the Western Cape Small and Medium-Sized Businesses. *International Conference on Cyber Warfare and Security*, 16, pp. 425-235. Academic Conferences Limited.

- Ncubukezi, T. and Mwansa, L. 2021. Best Practices Used by Businesses to Maintain Good Cyber Hygiene during Covid19 Pandemic. *Journal of Internet Technology and Secured Transactions*, 9(1), 714-721.
- Njoroge, GM 2020. *Human Factors Affecting Favourable Cybersecurity Culture-a Case of Small and Medium-sized Enterprises SMEs Providing Enterprise-Wide Information Systems Solutions in Nairobi City County in Kenya* (Doctoral dissertation, University of Nairobi).
- Richardson, M.D., Lemoine, P.A., Stephens, W.E. and Waller, R.E. 2020. Planning for Cyber Security in Schools: The Human Factor. *Educational Planning*, 27(2), 23-39.
- Sasse, M.A., Brostoff, S. and Weirich, D. 2001. Transforming the 'weakest link' —a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19(3), 122-131.
- Schneier, B. (2000). *Secrets and Lies: Security in a Digital World*. John Wiley & Sons.
- Solvere, O. 2021. *Cyber Attacks on Small Businesses Increase*. Available from <https://www.solveone.com/pages/cyber-attacks-on-small-businesses-increasing-in-2021/> [accessed on 16 September 2021].
- Tam, T., Rao, A., and Hall, J. 2021. The good, the bad and the missing: A Narrative review of cyber-security implications for Australian small businesses. *Computers & Security*, 102385.
- Teufel, S., Teufel, B., Aldabbas, M. and Nguyen, M. 2020. Cyber Security Canvas for SMEs. In *International Information Security Conference*, 20-33. Springer, Cham.
- Torten, R., Reaiche, C. and Boyle, S. 2018. The impact of security awareness on information technology professionals' behavior. *Computers & Security*, 79, 68-79.
- Turk, R.W. 2013. *Preparing a Cyber Security Workforce for the 21st Century*. Army War College Carlisle Barracks, PA.
- Wallace, S., Green, K., Johnson, C., Cooper, J., and Gilstrap, C. 2021. An Extended TOE Framework for Cybersecurity Adoption Decisions. *Communications of the Association for Information Systems*, 47 (2020), 51.
- Zimmermann, V. and Renaud, K. 2019. Moving from a 'human-as-problem' to a 'human-as-solution' cybersecurity mindset. *International Journal of Human-Computer Studies*, 131, 169-187.

Need for a Cyber Resilience Framework for Critical Space Infrastructure

Syed Shahzad, Li Qiao and Keith Joiner

School of Engineering and IT, Capability Systems Centre, University of New South Wales, Canberra., Australia

s.shahzad@adfa.edu.au

l.qiao@adfa.edu.au

k.joiner@adfa.edu.au

Abstract: The purpose of this paper is to introduce a case for a standardised comprehensive cyber resilience framework for Critical Space Infrastructure (CSI). Based on structure systematic review and meta-analyses, this paper outlines the needs of a risk-based framework. Space assets are fundamental components of critical national infrastructure (CNI), whose destruction significantly impacts many lives. Moreover, today's digitally connected space infrastructure is exposed to sophisticated and catastrophic cyber-attacks. This paper lays out the research gap to present the need for a comprehensive cyber resilience framework for CSI and future research and collaboration to understand the emergence of a new category of failures related to space-asset-reliance disruption risks.

Keywords: Cyber Resilience Framework for Critical Space Infrastructure, Critical Infrastructure, Cyber Security, Cyber Resilience, Resilience Framework

1. Introduction

Space assets are fundamental components of critical national infrastructure. They are an integral and crucial part of modern navigation, communication, weather, financial services, timing, defence, and science (Gheorghe et al., 2018). Space services enable society's vital functions, and their service offerings are difficult to substitute. The physical, cyber, geographical, and logical aspects of space infrastructure make it a perfect complex system of systems subject to cascading and escalating failures, increasing the impact of damage (Mureşan and Georgescu, 2017). Mureşan et al. (2016) state that space systems are a new type of critical infrastructure, not just a new category. Georgescu et al. (2019) argue that space assets should be described and accepted as critical national infrastructure. Although there has been much research in addressing space infrastructure operational security problems, little effort is made in developing a common means of assessing the resilience of the critical space infrastructure. A critical space infrastructure is a combination of space and ground assets whose disruption or failure would impact other critical infrastructure (Georgescu et al., 2018). A disproportionately small number of studies explicitly investigate the vulnerability of the critical space infrastructure to cyber-attack scenarios and how they could be prevented, mitigated, and made resilient. Few scholars have drawn on systematic research into the cyber resilience framework for critical space infrastructure. Therefore, there is an urgent need to address the resilience for space infrastructure security as a field. As the first step towards the cyber resilience framework, the current paper aims to understand the need for such a framework. Therefore, the overall motivation behind this work is to initiate a discussion for a comprehensive cyber resilience framework for critical space infrastructure. This work is the first known to systematically examine the research in cyber resilience for critical space infrastructure and calls for a comprehensive framework.

2. Background

Much of the research up to now has been limited and focused on one particular domain of space, which is operational security (OT). However, the unique threats for space infrastructure come from four separate attacks and weapons known as anti-satellite weapons, described as "*kinetic physical, non-kinetic physical, electronic, and cyber*" (Way, 2021). As most digitised critical infrastructure, space infrastructure is equipped with computers and networked hardware with connectivity for remote analytics, configuration, and upgrades, creating opportunities for cyber-attacks and cyber-defence. Typically, space infrastructure is divided into three main segments: outer space equipment, ground stations, and launch providers. These segments have unique vulnerabilities (Usman et al., 2020). The interdependencies between the three space segments lead to the proliferation of unique space cyber security risks and threats (Georgescu and Bucoveţchi, 2017). Despite the importance and known limitations of cyber security for space infrastructure, there has been little focus on cyber resilience for space assets and comprehensive guidance in the form of standards from the technology, ownership, and management perspectives.

This paper focuses on cyber-attacks and work published within the tenants of cyber resilience. The study's goal is to understand better the needs and the performance of the cyber resilience framework. Therefore, this paper motivates the need for such as cyber resilience framework and outline elements of an evaluating cyber resilience framework. To the authors' best knowledge, this study is unique in the discipline since this work is the first to examine the research and calls for establishing a need for a cyber resilience framework for critical space infrastructure. Since the topic crosses multiple domains, it consists of cyber, space and system engineering. In order to fully cover these disciplines' branches, a more advanced reporting system is required to overview the topic. The preferred reporting items for systematic reviews and meta-analyses reporting guidelines (Luhnen et al., 2018) is a comprehensive, explicit, and reproducible method, as documented guidelines by Durach et al. (2017).

The remainder of this paper is organised as follows. Section two presents the background of the study and motivates the need for a cyber resilience framework. Then, section three introduces the method to identify the need for such a framework. The second four discusses the finding, section five outlines the first steps towards a common resilience framework. Finally, Section six concludes the paper and section seven presents the future work.

3. Methodology

In order to assess the need for a cyber resilience framework for critical space infrastructure, the adopted method is the preferred reporting item for systematic reviews and meta-analyses (PRISMA), based on English-language peer-reviewed papers published in the last ten years. PRISMA is widely known as an instrument to qualitatively appraise a systematic review of the quality attributes evaluations (Luhnen et al., 2018). The search results are visually represented in Figure 1, following the PRISMA set structure (Moher et al., 2009). The selected search terms (keywords), including the alternate terms used, are listed in Table 1. The search criteria were applied to SCOPUS, Google Scholar, IEEE, and JSTOR databases, and a spreadsheet was developed to construct combinations of keywords to address all research questions. The articles between 2017 and 2021 were shortlisted whose abstracts, titles, or keywords included one of the search terms. Approximately 241,000 documents were published or accepted between 2017 and 2021 (up to and including mid-Oct 2021 at the time of writing) in cyber security, space, systems engineering and space engineering domains. Over 467 articles were finally selected and included in this literature review for screening. Forty-six per cent of the articles screened were excluded based on non-relevance. Another thirty per cent of articles were not retrieved because of record-duplication between databases. The final 115 most applicable and available articles were read carefully to answer, "What frameworks, standards and regulations are developed explicitly for cyber resilience of space infrastructure?".

Most of the literature used was published between 2019-2021, and 68% of the studies go back no more than seven years. USA, China, Romania, and Turkey have 40% contributions and pay attention to critical space infrastructure. There is a significant lack of peer-reviewed work in the area of cyber resilience for space infrastructure.

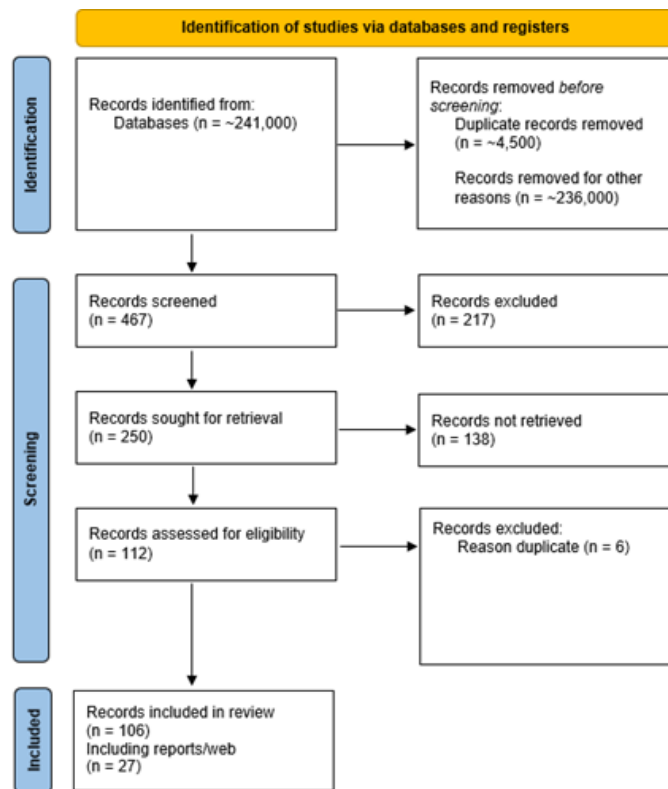


Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Table 1: Search keywords and combinations

Keyword	Condition	Keyword
space assets	or	cybersecurity
critical space infrastructure	or	standards
space systems cybersecurity	and	framework
space assets frameworks	and	cybersecurity
space assets policy	and	cybersecurity
space assets policy	or	cybersecurity
critical space infrastructure	or	cyber resiliency
Space infrastructure	and	cyber resiliency
Space infrastructure	and	cyber
space cyber frameworks	or	assessment
cybersecurity assessment frameworks	and	Space
critical space infrastructure	and	cyber attacks
Space infrastructure	and	governance
Space infrastructure	and	cyber attacks
Space infrastructure	and	kinetic attacks
Space infrastructure	and	attacks
Space infrastructure	and	current threats
cyber vulnerability assessment	or	space
attack vectors	or	critical assets
cyber security threats	and	SCADA or ICS or IoT
cyber security space	and	security models
critical infrastructure	and	attacker behaviour
cyber attack	and	motivation
system engineering	and	cybersecurity (cyber security)
system engineering	and	cyberresiliency (cyber resiliency)
system engineering	and	space
Space system engineering	and	security
system engineering frameworks	and	space
systems engineering	and	business
systems engineering	and	business process / management
systems engineering	and	business framework
systems engineering	and	business management
systems engineering	and	business process modeling
system engineering	and	risk management
regulatory framework	and	cybersecurity (cyber security)
regulatory framework	and	cyber resiliency/resilience
regulatory framework	and	space
business efficiency	and	space
business	and	regulatory framework
cybersecurity	and	risk assessment frameworks
cybersecurity	and	asset classification
cybersecurity	and	risk categorisation
cybersecurity	and	business impact
cybersecurity	and	risk assessment

4. Analysis and Discussion of Findings

The outer space assets are subjected to specific space phenomena, orbital debris, and the harshness of the outer space environment with temperatures and radiation (Mark, C.P. and Kamath, S., 2019). However, deliberate threats to critical space systems come from the state, rogue states, and non-state actors with various access levels to highly capable and affordable anti-satellite weaponry (La Bella, 2020). Egeli (2021) argues that the critical space infrastructure is under new anti-satellite attacks, and adversaries are increasingly developing new cyber threats, perhaps to achieve coordinated cyber-storming within 'grey-zone' tactics short of actual war (Austin, 2020). In addition, the kinetic weapons to destroy satellites in low Earth orbit have been shown to cause a large debris field that threatens other satellites and pollutes the space domain beyond recovery (Egeli, 2021). Bateman (2021) argues that the United States should avoid kinetic anti-satellite. Instead, the US should invest in developing and deploying non-kinetic weapons such as electronic or cyber-attacks against their adversaries' space assets.

In contrast to kinetic weapons, the impact of cyber-attacks in outer space is not well understood. Georgescu et al. (2019) found air-transport infrastructure was virtually dependent on space systems, and the space systems have a high probability of spontaneous malfunction as they operate in the most challenging environments. Johnson and Yepez (2011) analysed the Global Navigation Satellite Systems (GNSS) cyber threats. Their research intends to *provide an integrated, risk-based approach to the identification of attack scenarios that can help assess the resilience of safety cases to security threats.* They also investigated several threats to GNSS infrastructures from denial-of-service attacks on the ground-based infrastructure, data integrity, and vulnerabilities through insider attacks that could affect the services' accuracy, integrity, availability, and continuity. However, their focus was on the safety aspects and did not perform a fuller cyber risk assessment.

The growth in outer space activities is mainly driven by the commercial demand from new satellites for space transportation. This growth is called "Space 2.0", where *'out of new space thinking has come new technologies, new market entrants, new launcher systems, new ways of financing space ventures, new satellite architectures, efficient new small satellite designs, new types of ground antenna systems with electronic tracking, and market shifts toward networked services'* (Logue, T.J. and Pelton, J., 2019). The private sector plays a crucial role in the outer space activities where the public sector once was driving innovation and technology. In 2021, more than 50 countries will operate satellites in space (The Global Nature of Space Activities | Secure World, 2021). The new wave of smaller and more advanced satellites creates the danger of space crowding and signal interference. Safe operations of space assets are not only critical for the owners of those assets but global security and peace, as space assets monitor anti-terrorist activities, surveillance, and coordinate and disaster management (Bhattacharjee et al., 2018). Significant dependency on space infrastructure poses a critical yet under-recognised security predicament for critical infrastructure providers, policymakers, and governments.

Surprisingly, Australia, one of the wealthiest nations, has contributed little to space policy and space in general (Biddington, 2021). Georgescu et al. (2015) point out that governance in the space environment is complex since space has no clear jurisdictional boundaries that inform critical national infrastructure processes on Earth. For example, the European Commission states: *'Space infrastructure is critical infrastructure on which services that are essential to the smooth running of our societies and economies and our citizens' security depend. It must be protected, and that protection is a major issue for the EU which goes far beyond the individual interests of the satellite owners'* (EU., 2008). Therefore, each country currently takes responsibility to identify, designate and protect its specific infrastructure as critical national infrastructure, with no overarching regulatory authority.

Although space technology plays a critical role concerning commerce and security, it can also be seen as a threat to international peace (Rao et al., 2017). Space was the frontier for governments who were the primary sponsor of exorbitant research and development. Hence the International space law does not cover the space activities of today's major player, the private sector. By law, the private sector entities are to be under the supervision of the state of their jurisdiction, and their activities are to be assumed within the framework of relevant national laws and policies. *'There is no national legislation in many states on space-related matters, and many key issues relating to liability, technical safeguards, security, and innovation are not addressed adequately'* (Oltrogge et al., 2020). There is an urgent need for quite a few nations to develop legislation and standardise them with International Law. There has been a drive from the international community to develop "risk-based" frameworks for both public and private sectors (Daniels, and Paté-Cornell, 2017). Despite increasing space

technology advancements with rapid growth, international space lawmakers have not progressed beyond the initial five treaties and non-binding resolutions in the last fifty years.

Unfortunately, the UN treaties on outer space activities do not adequately facilitate commercial space activities (Rao et al., 2017). The current corpus of space law comprises treaties and agreements accepted and adopted by the total signatories. Such non-universal recognition and buy-in have seriously undermined the universal applicability of space law despite its adoption by unanimous approval in the UN General Assembly (Sachdeva, 2017). Oralova (2015) describes the *"Jus Cogens of Space Law"* as the fundamental principles that *'carry unanimous acceptance among the comity of nations as peremptory norms and command invariable adherence from states.'* *'Potential derogation from the principles of non-appropriation of outer space and of freedom of exploration and use of outer space, which is essential for the whole system of international space law is inconsistent with the natural state of outer space and can threaten international space and security'* (Oralova, 2015). These laws state as follows:

- Outer space as a province of humankind
- Freedom of access to all states for exploration and use
- State responsibility to humanity
- Prohibition on the placement of weapons in earth orbit
- Rescue and return of astronauts and space objects

The UNGA Resolution on Legal Principles recognised the risk of outer space weaponisation. However, it did not consider potential digital or, now as we know, cyber threats (UNOOSA, 2021), nor has the resolution been updated to reflect the new challenges. The UNGA authors only considered the possibilities relating to kinetic attacks, such as military activities. The North Atlantic Treaty Organisation (NATO) has facilitated discussions to address cyber related challenges and made series of recommendations (Tatar, U. 2020).

The increasing use of space systems for military reasons and several other conflicting issues make space assets vulnerable to attack. As per the United Nations Office for Outer Space Affairs (UNOOSA), *'State-Parties undertake not to place in orbit around the Earth any objects carrying nuclear weapons or any other kinds of weapons of mass destruction, install such weapons on celestial bodies, or station weapons in Outer Space in any other manner'* (UNOOS 2, 2021). The Secretary-General of the United Nations Group of Governmental Experts commissioned a study on transparency and confidence-building measures in outer space activities. The study concluded that *the 'World's growing dependence on Space-based systems and technologies and the information they provide requires collaborative efforts to address threats to the sustainability and security of Outer Space activities'* (Murthi, Gopalakrishnan, 2017). Somewhat in contrast to the research literature, governments and industry have begun to address cyber security for space infrastructure. For example, in a recent initiative, the White House released a Space Policy Directive (SPD5) entitled *"Cybersecurity Principals for Space Systems"* (Directive5, 2020) that establishes a broad set of guidelines for space companies in developing their cyber protection approaches. These guidelines include:

- Protection against unauthorised access to critical space vehicle functions;
- Physical protection measures designed to reduce the vulnerabilities of a space vehicle's command, control, and telemetry receiver systems;
- Protection against communications jamming and spoofing;
- Protection of ground systems, operational technology, and information processing systems through adopting deliberate cyber security best practices. This adoption should include practices aligned with the National Institute of Standards and Technology's cyber security framework to reduce the risk of malware infection and malicious access to systems, including from insider threats;
- Adoption of appropriate cyber security hygiene practices, physical security for automated information systems, and intrusion detection methodologies for system elements; and
- Management of supply chain risks that affect space systems' cyber security by tracking manufactured products, requiring sourcing from trusted suppliers; identifying counterfeit, fraudulent, and malicious equipment; and assessing other available risk mitigation measures.

These principles are defined as best practices, but they lack a formal framework and governance toolkit. MITRE has, however, released a recent paper, *"Cyber Best Practices for Small Satellite"*, to address some concerns raised by SPD5 (Visner and Kordella, 2020). A new Space Information Sharing and Analysis Centre (ISAC) has also recently been established *'to enable collaboration across the global space industry to enhance our ability to*

prepare for and respond to vulnerabilities, incidents, and threats; to disseminate timely and actionable information among member firms, and to serve as the primary communications channel for the sector with respect to this information' (Space ISAC. 2021). ISAC has developed a framework to describe threats in operations, business systems, provider, and communication segments. Besides that, a handful of private organisations have published some work, such as Aerospace Industries Association, published the first National Aerospace Standard NAS9924, citing it as "*Cyber Security Baseline*". Aerospace Industries Association developed baseline controls that focus on cyber security for the supply chain for aerospace and defence supply (Aerospace Industries Association, 2021)

Researchers have also become increasingly interested in anti-satellite cyber threats in recent years as their strategic, legal, and technical implications are not well-understood (Heinl, 2016). Scholars have argued the limitations of the '*non-binding doctrine*' of *jus ad Bellum* to cyber conflicts "Tallinn Manual" (Wallace and Jacobs, 2018) (Schmitt, 2017) (Norris, 2013). Therefore, the significant dependency on space infrastructure poses a critical yet under-recognised security predicament for critical infrastructure providers and policymakers. The threats for critical space infrastructure arise from distinct types of attacks and weapons known as anti-satellite (ASAT) weapons. These weapons are kinetic, non-kinetic, electronic, and cyber. If a significant cyber-attack happens, the resilience of commercial space infrastructure is likely to be fragmented, reduced, and may require expensive replacement, such that key-dependent terrestrial services would be unavailable for long durations. Therefore, a contemporary challenge is developing, mandating, and implementing cyber security and resiliency for a complex and substantial space infrastructure spanning different systems, jurisdictions, and legal frameworks. However, there is generally a lack of peer-reviewed research in cyber security for space infrastructure, and not much work has been done in space reference architecture, governance frameworks, and tools. Thus, the paper suggests a comprehensive cyber resilience framework is needed to regulate the expanding space industry.

5. Towards a Cyber Resilience Framework

This section outlines the needs for the future cyber resilience framework for critical space infrastructure. First, the future framework shall build on a set of risk assessment and management questions. The overarching approach for the suggested framework is based upon Kaplan and Garrick (1991) structured risk analysis employing the following three fundamental risk assessment questions: 'what can go wrong?', 'what is the likelihood?', and 'what are the consequences?' In addition, the future cyber resilience framework shall cover the next triplet of framework-specific questions:

1. What viable options are available?
2. What is the actual cost of risk?
3. What is the return on risk reduction investment?

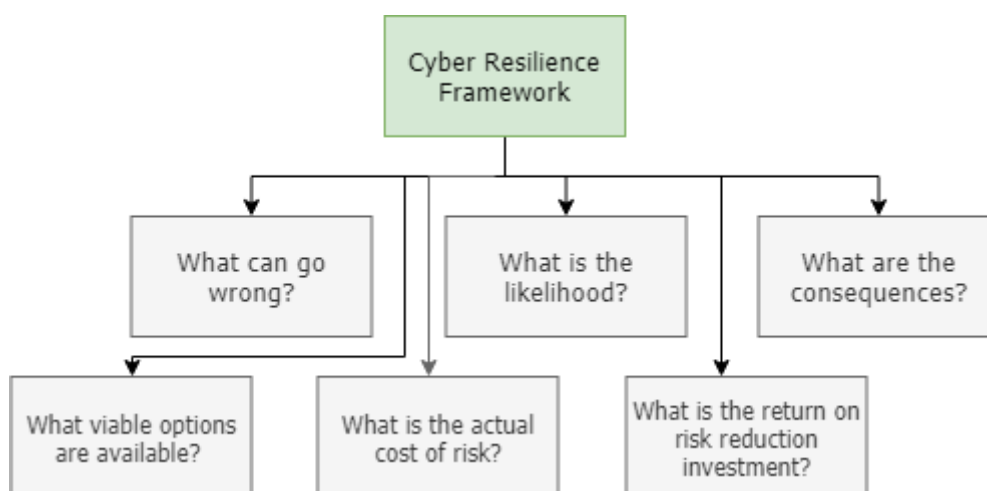


Figure 2: Framework Foundation based on Structured Risk Analysis

Second, the future framework shall be fair and consistent. It assures that the development and integration of a space system (and subsystems) hardware, software and supply chain fully incorporated cyber security and resilience needs appropriate to the contested information environment for which it is intended to operate. The framework employs attack surface management (ASM) for identification, inventory, classification, monitoring

and prioritisation of all digital assets of a space system. Our proposed framework builds on the above risk assessment and management questions, developing a discussion on the current literature on the subject, and compares and highlights the similarities and new possibilities.

Third, the future framework shall consider both tangible and intangible assets. Despite decades of research on cyber security and risk assessment, finding the critical assets (tangible and intangible), and more importantly, the value of “intangibles” has been less than satisfactory. Therefore, asset categorisation, risk identification, risk assessment and decision-making methods are essential for the future cyber security framework to provide business impact assessment.

Fourth, the future framework building shall develop a reliable cyber resilience framework and ontology encompassing many requirements. For example, the framework will need to be risk-based (cyber and business risks), requirements focused, and cost-driven (correct cost estimation for impact analysis). Further, the framework will need to incorporate security by design (cyber security and resilience into design, development, acquisition, operation, and sustainment), be resilient (protect, detect, react, and restore), be controls-based (select, implement, assess, and monitor) and where possible reduce the test and evaluation burden.

6. Conclusion

The critical space infrastructure (CSI) supports vital functions of society to operate its sensitive and crucial services from health, mobility, and security and peace. Despite their importance and increased cyber security risks, threats and attacks, the global community is yet to define, negotiate and mandate a treaty or instrument. At present, there are no agencies or governing bodies that monitor and restrict the use of satellites (Falco, 2018). The lack of overarching governance creates a vacuum in security design and operational cyber resiliency for the private sector that plays a vital role in the outer space activities where the public sector once was driving innovation and technology. Modern space systems are fundamentally dependent upon information technology and networked systems operating in the domain of cyberspace. These systems are inherently interconnected, either directly or indirectly, formatting a complex system of systems whose operations are interdependent and critical to the command, control, and performance of platforms, sensors, and other elements. Although the dependence on cyberspace and the networking of systems is a double-edged sword, enhancing our capabilities yet at the same time introducing numerous potential vulnerabilities. In response to the increasingly pervasive and adaptive threat environment, a consistent and fair cyber security and resilience assessment framework are critical to reduce the risk and provide a layered approach to ensure mission assurance.

7. Future Research

While the concept of a cyber security and resilience framework for space infrastructure is new, it certainly demands the attention of academics from all domains. Comparisons between the commercial space domain and banking and finance can be drawn, where risk criticality and cost are crucial for risk owners. At present, the space industry is like the High Seas and Antarctica, where no single regulatory or governance framework is applicable, instead of a combination of international treaties, commercial contractual frameworks, and national laws only relevant to specific objects and individuals. Remarkably few studies have been designed to study the commercial space assets cyber security and resiliency from a risk-based, cost-benefit approach. As critical space infrastructure is under new anti-satellite attacks and the private sector is increasingly growing, this literature review suggests developing and research a structured cyber-resilience framework for critical space infrastructure with:

- an underpinning cyber-vulnerability ontology for finding the critical digital assets,
- a means to value “intangibles”, and
- a governance framework for through-life secure and resilient space operations.

A disproportionately small number of studies explicitly investigate the vulnerability of the critical space infrastructure to cyber-attack scenarios and how they could be prevented, mitigated, and made resilient. Therefore, we suggest that coherent and efficient future collaborative research can be significantly aided by understanding the already proposed cyber resilience frameworks in other domains.

References

- Aerospace Industries Association. 2021. AIA Announces First Cyber Security Standard | Aerospace Industries Association. [online] Available at: <<https://www.aia-aerospace.org/news/aia-announces-first-cyber-security-standard>> [Accessed 13 October 2021].
- Austin, G. ed., 2020. National cyber emergencies: The return to civil defence. Routledge.
- Bateman, A., 2021. America Can Protect Its Satellites Without Kinetic Space Weapons - War on the Rocks. [online] war on the Rocks. Available at: <<https://warontherocks.com/2020/07/america-can-protect-its-satellites-without-kinetic-space-weapons/>> [Accessed 13 October 2021].
- Bhattacharjee, D., Aqeel, W., Bozkurt, I.N., Aguirre, A., Chandrasekaran, B., Godfrey, P.B., Laughlin, G., Maggs, B. and Singla, A., 2018, November. Gearing up for the 21st century space race. In Proceedings of the 17th ACM Workshop on Hot Topics in Networks (pp. 113-119).
- Biddington, B., 2021. Is Australia Really Lost in Space?. *Space Policy*, 57, p.101431.
- Daniels, M.P. and Paté-Cornell, M.E., 2017. Risk-based comparison of consolidated and distributed satellite systems. *IEEE Transactions on Engineering Management*, 64(3), pp.301-315.
- Directive5, 2020. Cybersecurity Principles for Space Systems. National Space Policy.
- Egeli, S., 2021. Space-to-Space Warfare and Proximity Operations: The Impact on Nuclear Command, Control, and Communications and Strategic Stability. *Journal for Peace and Nuclear Disarmament*, 4(1), pp.116-140.
- EU., 2008. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data, COM/2020/66 final. [online] Op.europa.eu. Available at: <<https://op.europa.eu/en/publication-detail/-/publication/ac9cd214-53c6-11ea-aece-01aa75ed71a1/language-en>> [Accessed 13 October 2021].
- Falco, G., 2018. The vacuum of space cyber security. In 2018 AIAA SPACE and Astronautics Forum and Exposition (p. 5275).
- Georgescu, A. and Bucovetchi, O., 2017, May. A generic flow based model for understanding critical infrastructure dependency on space systems. In Proceedings of the 29th international business information management association conference (pp. 3-4).
- Georgescu, A., Bucovetchi, O. and Tatar, U., 2018. Space systems as critical infrastructures. *FAIMA Business & Management Journal*, 6(1), pp.24-34.
- Georgescu, A., Gheorghe, A.V., Piso, M.I. and Katina, P.F., 2019. Critical Space infrastructures: risk, resilience and complexity (Vol. 36). Springer.
- Gheorghe, A.V. and Vamanu, D.V., 2007. Risk and vulnerability games. The anti-satellite weaponry (ASAT). *International journal of critical infrastructures*, 3(3-4), pp.457-470.
- Gheorghe, A.V., Georgescu, A., Bucovetchi, O., Lazăr, M. and Scarlat, C., 2018. New Dimensions for a Challenging Security Environment: Growing Exposure to Critical Space Infrastructure Disruption Risk. *International Journal of Disaster Risk Science*, 9(4), pp.555-560.
- Heinl, C.H., 2016. The potential military impact of emerging technologies in the Asia-Pacific region: A focus on cyber capabilities. In *Emerging Critical Technologies and Security in the Asia-Pacific* (pp. 123-137). Palgrave Macmillan, London.
- Kaplan, S., 1991. The general theory of quantitative risk assessment. In *Risk-Based Decision Making in Water Resources V* (pp. 11-39). ASCE.
- La Bella, J., 2020. Star Wars: Attack of the Anti-Satellite Weapons in Anticipatory Self-Defense. *U. Pac. L. Rev.*, 52, p.733.
- Logue, T.J. and Pelton, J., 2019. Overview of commercial small satellite systems in the "New Space" age. *Handbook of Small Satellites: Technology, Design, Manufacture, Applications, Economics and Regulation*, pp.1-18.
- Luhnen, M., Waffenschmidt, S., Schwalm, A., Gerber-Grote, A. and Hanke, G., 2018. Quality Assessment of Systematic Reviews of Health Economic Evaluations: Pitfalls with the Application of the PRISMA Statement. *Comment on Quang et al.(Sys Rev Pharm. 2017; 8 (1): 52-61). Systematic Reviews in Pharmacy*, 9(1), pp.83-84.
- Mark, C.P. and Kamath, S., 2019. Review of active space debris removal methods. *Space Policy*, 47, pp.194-206.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and Prisma Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7), p.e1000097.
- Muresan L., A. Georgescu, I. Jivănescu, S. Popa, and S. Arseni. 2016. Charting critical energy infrastructure dependencies on spaces Critical energy infrastructure protection and cyber security policies, ed. M.H. Casin, and G. Gluschke, 177–192. 2016. Istanbul, Turkey: Hazar Strateji Enstitu'su".
- Murthi, K.S. and Gopalakrishnan, V., 2017. Trends in Outer Space Activities—Legal and Policy Challenges. In *Recent Developments in Space Law* (pp. 27-42). Springer, Singapore.
- Norris, M.J., 2013. The Law of Attack in Cyberspace: Considering the Tallinn Manual's Definition of 'Attack' in the Digital Battlespace. *Inquiries Journal*, 5(10).
- Oltrogge, D.L. and Christensen, I.A., 2020. Space governance in the new space era. *Journal of Space Safety Engineering*, 7(3), pp.432-438.
- Oralova, Y., 2015. Jus cogens norms in international space law. *Mediterranean Journal of Social Sciences*, 6(6), p.421.
- Rao, R.V., Gopalakrishnan, V. and Abhijeet, K. eds., 2017. *Recent Developments in Space Law: Opportunities & Challenges*. Sachdeva, G.S., 2017. Select Tenets of Space Law as Jus Cogen. In *Recent Developments in Space Law* (pp. 7-26). Springer, Singapore.

- Schmitt, M.N. ed., 2017. Tallinn manual 2.0 on the international law applicable to cyber operations. Cambridge University Press.
- Space ISAC. 2021. Space ISAC - Space Information Sharing and Analysis Center. [online] Available at: <<https://s-isac.org/>> [Accessed 13 October 2021].
- Swfound.org. 2021. The Global Nature of Space Activities | Secure World. [online] Available at: <<https://swfound.org/space-sustainability-101/the-global-nature-of-space-activities/>> [Accessed 13 October 2021].
- Tatar, U., 2020. Critical Space Infrastructures: Perspectives and a Critical Review. Space Infrastructures: From Risk to Resilience Governance, 57, p.3.
- UNOOS 2., 2021. ST/SPACE/61/Rev.2: International Space Law: United Nations Instruments. [online] Unoosa.org. Available at: <https://www.unoosa.org/oosa/oosadoc/data/documents/2017/stspace/stspace61rev.2_0.html> [Accessed 13 October 2021].
- UNOOSA. 2021. Legal Principles. [online] Available at: <<https://www.unoosa.org/oosa/en/ourwork/spacelaw/principles/legal-principles.html>> [Accessed 13 October 2021].
- Usman, M., Qaraqe, M., Asghar, M.R. and Shafique Ansari, I., 2020. Mitigating distributed denial of service attacks in satellite networks. Transactions on Emerging Telecommunications Technologies, 31(6), p.e3936.
- Visner, S.S. and Kordella, S., 2020. Cyber Best Practices for Small Satellites. In ASCEND 2020 (p. 4013).
- Wallace, DA and Jacobs, C.W., 2018. Conflict Classification and Cyber Operations: Gaps, Ambiguities and Fault Lines. U. Pa. J. Int'l L., 40, p.643.
- Way, T., 2021. Counterspace Weapons 101 - Aerospace Security. [online] Aerospace Security. Available at: <<https://aerospace.csis.org/aerospace101/counterspace-weapons-101>> [Accessed 13 October 2021].

Ransomware Detection using Process Memory

Avinash Singh¹, Richard Adeyemi Ikuesan² and Hein Venter¹

¹University of Pretoria, South Africa

²Community College Qatar, Doha, Qatar

asingh@cs.up.ac.za

richard.ikuesan@ccq.edu.qa

hventer@cs.up.ac.za

Abstract: Ransomware attacks have increased significantly in recent years, causing great destruction and damage to critical systems and business operations. Attackers are unfailingly finding innovative ways to bypass detection mechanisms, which encouraged the adoption of artificial intelligence. However, most research summarizes the general features of AI and induces many false positives, as the behavior of ransomware constantly differs to bypass detection. Focusing on the key indicating features of ransomware becomes vital as this guides the investigator to the inner workings and main function of ransomware itself. By utilizing access privileges in process memory, the main function of the ransomware can be detected more easily and accurately. Furthermore, new signatures and fingerprints of ransomware families can be identified to classify novel ransomware attacks correctly. The current research used the process memory access privileges of the different memory regions of the behavior of an executable to quickly determine its intent before serious harm can occur. To achieve this aim, several well-known machine learning algorithms were explored with an accuracy range of 81.38% – 96.28%. The study thus confirms the feasibility of utilizing process memory as a detection mechanism for ransomware.

Keywords: Ransomware detection, process memory, memory-based malware analysis, machine learning, Cuckoo Sandbox

1. Introduction

Detecting ransomware attacks has always been a significant problem for organizations (Singh, Ikuesan & Venter, 2019a,b). Ransomware is the most dominant form of malware that causes major disruptions to daily business operations (Hampton, Baig & Zeadally, 2018; Newman, 2018; Sienko, 2021). The primary objective of ransomware is to encrypt sensitive files and hold the decryption of those files as a ransom for large amounts of money (Singh *et al.*, 2019a). In most instances, ransomware is delivered through major exploits or social engineering. According to the State of Ransomware 2021 report (Sophos, 2021), 54% of organizations that were hit by ransomware had their data encrypted. However, it is estimated that the average ransom payment made by mid-sized organizations was approximately USD 170,404 (Sophos, 2021). The report (Sophos, 2021) also calculated the average cost to recover from a ransomware attack to be around USD 1.8 million. Despite such attacks being a growing concern, the good news is that they have decreased slightly since 2017. This was the time when WannaCry wreaked havoc over the internet with more than 54% of organizations hit, but the number of attacks gradually decreased to around 51% in 2020 and 37% in 2021 (Sophos, 2021).

An alarming fact in the Sophos (2021) report is that 47% of respondents expected ransomware to hit their organization because “Ransomware attacks are increasingly hard to stop due to their sophistication”. Even though recovery from ransomware attacks has increased through implicit and immutable backups, there have been significant improvements in methods to detect ransomware attacks. Unfortunately, these methods are easily bypassed by sophisticated ransomware, which leaves security researchers in disarray to find effective detection methods. For instance, to achieve speed and efficiency in detection, anti-virus programs use static and signature-based detection methods. Static analysis is limited to known signatures and snippets of codes. Furthermore, the concept of signature-based detection suggests that only known attributes and unique behaviour can be identified. Thus, both signature-based and static analysis are not efficient against novel ransomware, making these methods inefficient in detecting ransomware attacks. This is due to ransomware being more behavioural based. Since ransomware utilizes security features like encryption, it is difficult to detect new never-seen-before ransomware (Bromium Labs, Kotov & Rajpal, 2014). To solve these issues, researchers have adopted Artificial Intelligence (AI) to detect ransomware attacks more accurately. However, there are many ways AI can be used to detect ransomware, and many researchers make use of automated feature extraction. Certain aspects may sometimes be overlooked, as the features are sourced from techniques and patterns/trends seen in the dataset that the researchers have – thus making it less accurate or slightly biased. It is important for experts or through rigorous experimentation to identify and hand select features that would provide the most insight and that cannot be mocked or influenced. In this paper, process memory was examined to determine

how accurately it can detect ransomware. Process memory is behaviour-based and one of the core functions of ransomware to encrypt files (Subedi, Budhathoki & Dasgupta, 2018).

Going forward, Section 2 provides the necessary background and related works, while Section 3 details the proposed method and results of ransomware detection. Section 4 evaluates the performance of the proposed method. The conclusion and future works are presented in Section 5.

2. Related Work

Over the years, significant advances have been made in ransomware detection, especially after the devastation that WannaCry caused in 2017 (Adamov & Carlsson, 2017; Berrueta, Morato, Magana, *et al.*, 2020; Fernando, Komninos & Chen, 2020; Molina, Torabi, Srieddine, *et al.*, 2021; Singh *et al.*, 2019a). Although researchers explored avenues for detection such as static and dynamic information, ransomware has managed to evade static analysis (Subedi *et al.*, 2018). Due to this limitation, typically active behavioural-based detection techniques are used, such as API calls, registry information, opcodes, manual analysis, and dissection (Molina *et al.*, 2021). Manual analysis is quite time-consuming and has led to several sandbox environments with analysis packages like Cuckoo Sandbox (Guarnieri, 2014), which is widely used by security researchers. Machine Learning (ML) has been widely utilized for the detection of malware as well as ransomware (Dang, Di Troia & Stamp, 2021; Dua & Du, 2016; Singh & Singh, 2021). For example, BlackBerry Cylance Anti-Virus leverages ML techniques to provide more robust protection against unknown attacks (BlackBerry, 2021). This causes challenges with detection rates and the way detection happens, as researchers were able to bypass Cylance protection (SCMagazine, 2019). There is a plethora of research that uses various features for ransomware detection, for example, the work by Ashraf *et al.* (2019) explores static and dynamic information for detection such as API calls, registry, file operations, strings, DLLs. Using these features, Ashraf *et al.* (2019) managed to obtain an accuracy of 92% for a support vector machine and 91% for a random forest using the dynamic information. The work by Fernando, Komninos, and Chen (2020) explored many ML algorithms that produced accuracies of between 33.1% and 96.4%, with the lowest being an SVM and the highest being a multi-layer perceptron with 10 hidden layers.

Findings in Arabo *et al.* (2017) explored malware detection using memory analysis of API calls, DLLs, process handles, network, code injection, and privilege as features in their dataset. The results obtained ranged from 95.94% to 98.50%, with the lowest being a random forest and the highest an SVM. However, the results for the random forest had a recall of 87.68%, which means it did not perform well in prediction. The SVM classifier on the other hand performed well and gave a low false-positive rate and recall. The study considered weighted features and had a large dataset of 3468 records. Similarly, Grégio *et al.* (2013) explored memory writes for malware classification and code-reuse identification. Clustering techniques were explored with a precision value of 84.3% using pre- and inter-clustering. To the best of the Author's knowledge, no research has been done using process memory specifically, rather process behaviour as seen in the work by Arabo *et al.* (2017). Their work involved collecting computer metrics such as the CPU, RAM, and disk usage, as well as the API calls and files open. The results of Arabo *et al.* (2017) were extremely poor and ranged from 52.85% for a neural network to 75.01% for a random forest. Existing literature (Dang *et al.*, 2021; Molina *et al.*, 2021; Rughani, 2017; Shah & Issac, 2018) shows there has been great reliance on the following ML algorithms: neural network; k-nearest neighbours; decision tree; random forest; Naïve Bayes; SVM; and boosted learners.

Exploring process memory specifically has been overlooked in the research space, since more data-rich features were available. However, these data-rich features often pose a degree of bias or allow attackers to easily bypass detection methods through deterrence mechanisms and obfuscation (Bashari Rad, Masrom & Ibrahim, 2012; Grégio *et al.*, 2013). Process memory can help identify the inner workings of any executable and with the power of machine learning, this can become simpler thereby minimizing manual analysis of large amounts of data. The current research paper aims to leverage process memory to distinguish ransomware more accurately from benignware. It also aims towards active detection without the need of utilizing a sandboxed environment.

3. Ransomware Detection using Process Memory

The proposed method of ransomware detection involved exploring the different memory regions that occur in process memory. The overall process employed in this research is modelled and shown in Figure 1. The process model starts with capturing the process's memory using the Cuckoo Sandbox (Guarnieri, 2014). Thereafter, Cuckoo runs a default reporting module that analyses the collected memory dump and produces a raw report

with each memory region from process memory along with metadata. The access privileges that are occurring in each memory region are captured from each region. These privileges include Read (**r**), Read/Write (**rw**), Read/Execute (**rx**), Read/Write/Copy (**rwc**), Read/Write/Execute (**rwX**) and Read/Write/Execute/Copy (**rwxc**). The privileges determine what operations and what types of data are stored in memory to help find patterns in subroutines that can be attributed to specific operations. For example, if a segment within the memory region has an access privilege of **rx** and the type is **code**, this could be a subroutine involved with reading a file and then performing some computation on it. However, the privilege levels include several aspects, and one of the most interesting and rarely seen privileges is **rwxc**. Using this highest privilege level could mean that the executable can persist, extract, and execute code in memory. On the other hand, Read/Write/Execute (**rwX**) could indicate a dynamic subroutine-like encryption cipher. Process memory consists of many segments within a memory region and therefore the approach adopted in this research is a summation of each segment's privileges per memory region. Once the data has been collected, the next phase is to conduct the Machine Learning (ML) exploratory phase, which involves exploring various ML techniques for classification. The techniques this study explored are classifiers such as decision tree, random forest, gradient boost, XGBoost, tree ensemble, neural network, Naïve Bayes, and support vector machine (Fernando *et al.*, 2020). These algorithms were tested to see which provides the best accuracy in detecting ransomware and produces the least number of false negatives.

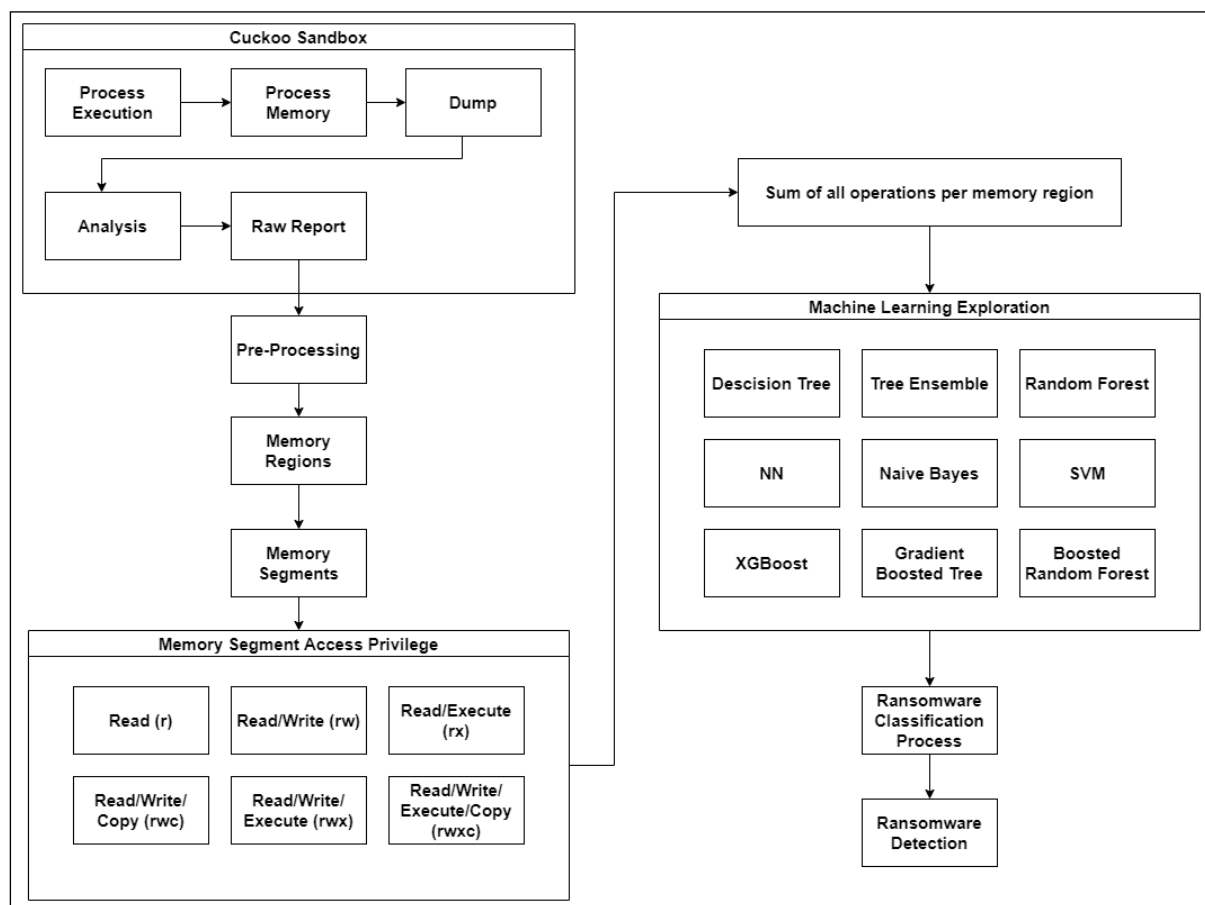


Figure 1: Ransomware Detection Process Model using Processor Memory

The subsections below discuss how the data was collected and pre-processed for the machine learning exploration phase. The results from the machine learning phase are also discussed.

3.1 Dataset

The most important aspect of machine learning is the feature engineering. Having enough information that represents all classes is extremely important to ensure a fair result in the learning phase. For this purpose, two classes were chosen for detection, namely benign and malicious.

3.1.1 Data Acquisition and Extraction Process

Acquiring the data involved a major process, which had to be done manually as there are no existing datasets that comprise process memory. Benign samples were obtained from various sources, with most originating from Portable Apps (Haller, 2021). A total of 354 benign samples were collected across different categories of executables. The categories explored were accessibility, development, education, games, image software, music/video software, office, security, utilities, internet, encryption, and general-purpose. These benign samples were taken at random from each category.

Malicious samples were sourced mainly from Malware Bazaar (Abuse.ch, 2020) and theZoo (Yuval Nativ, 2015). A total of 117 ransomware samples were sourced from over 70 categories of ransomware families. A snippet of the breakdown of each benign sample and malicious ransomware category is given in Table 1. The full dataset can be found at <https://github.com/ICFL-UP/RDPM>.

Table 1: Summary of samples in each category

Benign		Malicious					
Category	Count	Category	Count	Category	Count	Category	Count
accessibility	7	#Other	5	Clay	1	DarkSide	3
development	20	4444	1	Clop	3	DeathHiddenTear	1
education	11	ArkeiStealer	4	CobaltStrike	1	Demonware	3
games	20	Avaddon	3	Conti	3	Dharma	3
graphics_pictures	32	BB	1	CryLock	3	DoejoCrypt	3
internet	48	Babuk	4	Crypt888	1	Dridex	1
music_video	33	Balacclava	4	CryptoDarkRubix	1	EKANS	1
office	29	BlackKingdom	2	CryptoLocker	1	ESCAL	3
security	17	BlackRose	1	Cryptowall	1	Egregor	3
utilities	137	Cerber	1	Cuba	1	Encrpt3d	1

Cuckoo Sandbox (Guarnieri, 2014) was used to generate the process memory for each sample. Each sample was analysed with the default Cuckoo settings and the process memory was dumped and analysed with Cuckoo reporting modules to extract the process memory. The report included the regions, segments as well as the access privileges and type of content stored, start and end address, state, offset, and size. A snippet of the data obtained from Cuckoo can be seen in Figure 2. The current study explored the protect field, which contains the access privilege. It was observed to be suitable, as the other attributes remained constant or were dependant on how the operating system assigned memory (start/end addresses, and the offset).

```

"procmemory": [
  {
    "regions": [
      {
        "protect": "rw",
        "end": "0x00020000",
        "addr": "0x00010000",
        "state": 4096,
        "offset": 24,
        "type": 262144,
        "size": 65536
      }
    ]
  }
],

```

Figure 2: Process memory data from Cuckoo

3.1.2 Data Pre-Processing and Metrics

Each sample's memory region constituted a single record in the dataset. Some samples had multiple memory regions, while some only had one. Each memory region segment's access privileges from the Cuckoo reports were summed up for each privilege, thus resulting in the dataset as shown in Table 2. The dataset contained 476 benign records and 461 malicious records for a total dataset of 937 records and a total of 80 categories.

Table 2: Dataset sample

R	RW	RX	RWC	RWX	RWXC	LABEL	CATEGORY
367	307	117	84	70	0	B	Utilities
107	96	43	31	5	0	B	Utilities
62	59	18	21	49	0	B	Utilities
13	13	4	5	4	0	M	Cryptowall
124	94	54	32	6	0	M	DarkSide
239	289	79	104	98	22	M	DeathHiddenTear

3.2 Machine Learning Exploration Phase

Several classifiers (machine learning techniques) were explored with a training/testing ratio of 80/20 as the de facto standard (Ashraf *et al.*, 2019; Fernando *et al.*, 2020). Stratified sampling was used to ensure that the split was not biased or uneven in terms of the label distribution. Each algorithm was evaluated using the Knime Analytics Platform (Knime, 2021) on a Microsoft Windows machine. Metrics such as True Positive (TP), False Positive (FP), True Negative (TN), Recall, Precision, Sensitivity, Specificity, F-Measure, and Accuracy, were used to evaluate the result. The details of the explored algorithms are presented in the subsections that follow.

3.2.1 Decision Tree (DT)

A Decision Tree is one of the most commonly used classification algorithms in the security domain (Berrueta, Morato, Magana, *et al.*, 2019; Molina *et al.*, 2021). Trees provide a rule-based approach to classification, and since security relies on rules, this is often the preferred approach (Adeyemi, Razak, Salleh, *et al.*, 2016; Ernsberger, Ikuesan, Venter, *et al.*, 2017; Mohlala, Ikuesan & Venter, 2018). The parameters used for the DT comprised the quality measure that integrates the Gini Index, with no pruning, and a minimum number of five records per node. An accuracy of 93.62% (see Table 3) was observed. The number of incorrect classifications was 6:6 for benign:malicious samples respectively. This provides a fair misclassification. The Cohen's Kappa score of 0.8723 implies that although the DT model performed relatively better in classification. However, it is inefficient for detecting ransomware as there would be a 6.38% chance it would go undetected.

Table 3: Decision Tree Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	90	6	86	6	0.9375	0.9375	0.9375	0.9347	0.9375	
M	86	6	90	6	0.9347	0.9347	0.9347	0.9375	0.9347	
Overall										0.9362

3.2.2 Tree Ensemble (TE)

Ensemble learning is a technique that leverages concrete finite sets of models and evaluates them to find the best-performing model. Tree Ensemble is a model that builds multiple tree models and then evaluates each to find the best one. For the TE, the information gain ratio split criteria were utilized with the number of models set to 100 with square root attribute sampling. Increasing the number of models did not affect the accuracy, and a test was done with 10 models, which only affected the accuracy by 0.22%. Changing the split criteria and the sampling attributes had little to no effect on the accuracy of 95.74% (see Table 4). This is an improvement over the DT classifier, thereby providing better classification and a decrease in the confusion matrix of just two misclassifications for each label.

Table 4. Tree Ensemble Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	90	2	90	6	0.9375	0.9783	0.9375	0.9783	0.9574	
M	90	6	90	2	0.9783	0.9375	0.9783	0.9375	0.9574	
Overall										0.9574

3.2.3 Random Forest (RF)

Unlike a Tree Ensemble (TE), a Random Forest builds multiple tree structures within one model building a forest of trees. This classification algorithm is very useful for a complex hierarchy of rules and is particularly used often in the security domain for classification. While the RF did perform well it is slightly less accurate than the TE. This can be due to the dataset not having too many patterns from the different occurrences of the data points, resulting in a rather small forest, achieving a similar result as the TE. An RF and TE are very similar from an

algorithm point of view, therefore, having accuracy in the same range is expected. The accuracy statistics of the RF are shown in Table 5.

Table 5: Random Forest Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	89	2	90	7	0.9271	0.9780	0.9271	0.9783	0.9519	
M	90	7	89	2	0.9783	0.9278	0.9783	0.9271	0.9524	
Overall										0.9521

3.2.4 Gradient-Boosted Tree (GBT)

A Gradient-Boosted Tree is an ensemble learning where the notion of combining previous models would increase prediction. Hence, the goal is to minimize the gradient of the prediction error. Based on the results in Table 8, the accuracy is 94.68% with a false-positive misclassification of only four records. In comparison, DT had 6, whereas TE and RF had two, which means that GBT does not perform better. This is also verified by the accuracy score is lower than the other classifiers. Generally, it is said that gradient boosting gives a higher performance than a random forest, due to the optimization technique of reducing the prediction error.

Table 6: Gradient-Boosted Tree Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	90	4	88	6	0.9375	0.9574	0.9375	0.9565	0.9474	
M	88	6	90	4	0.9565	0.9362	0.9565	0.9375	0.9462	
Overall										0.9468

3.2.5 XGBoost

The XGBoost is a gradient boost algorithm with additional features that allow for better performance and accuracy. These features include a robust penalty function on the tree models, shrinking of leaf nodes, and the ability to deal with large data. While GBT did not perform well (see Table 7), the results of the XGBoost algorithm delivered better results, with a 96.28% accuracy and a relatively high precision and F-measure score further validating the accuracy results obtained. The XGBoost algorithm was configured with the softprob option to give the probability for each data point. A boost of 1000 rounds improved the accuracy.

Table 7: XGBoost Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	93	4	88	3	0.9688	0.9588	0.9688	0.9565	0.9637	
M	88	3	93	4	0.9565	0.9670	0.9565	0.9688	0.9617	
Overall										0.9628

3.2.6 Naïve Bayes (NB)

The Naïve Bayes algorithm tends to perform well, which results in higher accuracy than other algorithms. This is because the NB algorithm treats features individually, and these features are not related to each other. However, in this case, it performed poorly, which can be attributed partly to the smaller feature space and the fact that the features are related to one another. A Cohen's Kappa score of 0.6288 was obtained. The relatively low output indicates that this classification is not reliable. An accuracy score of 81.38% (see Table 6) further confirms that the NB algorithm is not efficient in detecting ransomware.

Table 8: Naive Bayes Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	71	10	82	25	0.7396	0.8765	0.7396	0.8913	0.8023	
M	82	25	71	10	0.8913	0.7664	0.8913	0.7396	0.8241	
Overall										0.8138

3.2.7 Support Vector Machine (SVM)

The Support Vector Machine algorithm was tested given that SVM works well with small datasets such as the one used in the current study. The SVM performs classification based on hyperplanes so that each label can be isolated from the other. However, the SVM algorithm did not perform well with this dataset, and it achieved an accuracy of 85.64% (see Table 10). A test was done using the configuration of a polynomial kernel where the

power, bias, gamma, and penalty were set to 1, which resulted in an accuracy of 52.45%. The final configuration (see the results in Table 10) used a radial basis function kernel with a sigma of 0.1.

Table 9: Support Vector Machine Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	94	25	67	2	0.9792	0.7899	0.9792	0.7283	0.8744	
M	67	2	94	25	0.7283	0.9710	0.7283	0.9792	0.8323	
Overall										0.8564

3.2.8 Neural Network (NN)

To compare a simple supervised learning classification to deep learning algorithms, a Neural Network was also explored. Resilient Propagation Multi-Layer Perceptron (RProp MLP) as defined by Riedmiller and Braun (1993) was used instead of a normal NN, because RProp MLP adjusts the weights of the network based on the behaviour of the error function. Since the data could not be used in its raw form, further pre-processing was needed for the NN training and all the features were normalised to values between 0 and 1. Different configurations of the parameters were explored, and the best performing parameters were 10000 iterations, two hidden layers, and 15 hidden neurons per layer. This resulted in an accuracy of 93.62% (see Table 11), which is on par with the tree classifiers.

Table 10: Neural Network Results

Label	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy
B	120	8	144	10	0.9231	0.9375	0.9231	0.9474	0.9302	
M	144	10	120	8	0.9474	0.9351	0.9474	0.9231	0.9412	
Overall										0.9362

4. Discussion

Overall, process memory gave a good measure for ransomware detection as most ML algorithms gave an accuracy of 93-96% (see Table 12). The Neural Network classifier also provided a relatively good performance. However, further exploration of other deep learning classifiers was discouraged in this study. A deep learning algorithm requires a relatively larger dataset, whereas the dataset used in this study was insufficient for typical deep learning algorithms to make good predictions. For instance, NN needs a larger dataset to generate a more accurate classification. Getting more data is a difficult process, as samples need to be manually run and a Cuckoo report can easily extend over 1 GB for each ransomware sample. Expanding on the dataset would also need more benign samples to maintain the class balance, in which case benign samples do not generate much process memory information – as opposed to ransomware. Machine learning algorithms suggested in (Berrueta *et al.*, 2019; Molina *et al.*, 2021; Rughani, 2017; Vinayakumar, Soman, Velan, *et al.*, 2017) obtained accuracy scores that ranged between 88-98%, which are good results. In comparison, the proposed research achieved the highest accuracy of 96.28% using the XGBoost algorithm. The limitation of the current research is that the dataset did not contain significantly larger training data. This can be attributed to the lack of resources and the absence of timeous acquisition, processing, and extraction of the process memory. Furthermore, the region of memory considered in this study is relatively measurable and logically feasible to provide a measure of dissimilarity between benignware and ransomware. The exploration of these memory-rich features presents a novel approach to ransomware investigation, consequently, improving ransomware mitigation when used in digital forensic readiness.

Table 11: Summary of Accuracy Statistics

TECHNIQUE	DT	TE	RF	NB	XG	GB	SVM	NN
Accuracy (%)	93.62	95.74	95.21	81.38	96.28	94.68	85.64	93.62

To further evaluate the degree of reliability, the area under the Receiver Operating Characteristic (ROC) curve – AUC – was explored. The ROC measures binary classifiers on their discriminative threshold based on the ratio of a true positive rate to a false positive rate. The AUC provides a basis for an unbiased evaluation of the performance of a classifier relative to the baseline (random) classifier. The AUC typically ranges from 0 to 1, with 0 being the worst classifier performance. Figure 3 shows the ROC curve for the classifiers explored in this study, with the AUC for each classifier shown in parenthesis in the legend. Looking at Figure 3, the worst-performing model in terms of the AUC output is the NN. Overall, the best performing model in terms of accuracy and AUC

is XGBoost (XG), with an AUC of 0.992 and an accuracy of 96.28%. Similarly, the performance of RF and TE respectively rivals XG. This echoes the observed accuracy in Tables 4 and 5, where both classifiers demonstrated similar performance.

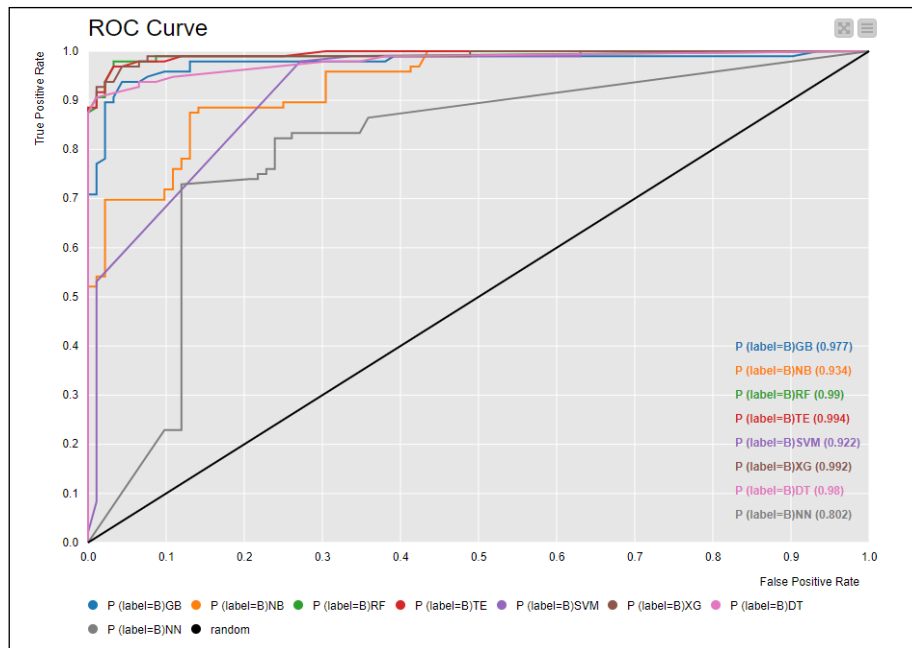


Figure 3: Area Under the ROC Curve of the Explored Classifiers

5. Conclusion and Future Work

In this paper, process memory was utilized for ransomware detection with the dataset comprising 937 records from 117 ransomware and 354 benign samples. The study observed, by leveraging several machine learning algorithms, that process memory does provide very meaningful information to assist with ransomware detection, and accuracy scores of between 93.62 and 96.28% were obtained, with the best performing model being XGBoost. Since static analysis often does not pick up ransomware attacks and behavioural analysis can be time-consuming, monitoring process memory is the most effective solution to actively detect ransomware attacks. Once process memory has been allocated, the access levels of the memory can provide meaningful information about the inner workings of the process. Data within the memory regions were not explored, as attackers would then make use of encryption to disguise their activity. Furthermore, the permissions of the access level cannot be hidden or encrypted, as this is done on an OS level. However, exploring the access level of a memory region is not sufficient for ransomware detection, therefore, future work would involve looking at the patterns in which these access levels occur for a more accurate and distinguishable mechanism for detection. Exploring the type of memory segments for each privilege can further provide more useful information to be used in prediction models. The paper in hand contributes to the security community by finding more novel ways for ransomware detection through automated behavioural analysis.

References

- Abuse.ch. 2020. *Malware Bazaar*. [Online], Available: <https://bazaar.abuse.ch/> [2021, June 11].
- Adamov, A. & Carlsson, A. 2017. The state of ransomware. Trends and mitigation techniques. in *Proceedings of 2017 IEEE East-West Design and Test Symposium, EWDTS 2017*.
- Adeyemi, I.R., Razak, S.A., Salleh, M. & Venter, H.S. 2016. Observing consistency in online communication patterns for user re-identification. *PLoS ONE*. 11(12):1–27.
- Ashraf, A., Aziz, A., Zahoora, U. & Khan, A. 2019. Ransomware Analysis using Feature Engineering and Deep Neural Networks. 1–15. [Online], Available: <http://arxiv.org/abs/1910.00286>.
- Bashari Rad, B., Masrom, M. & Ibrahim, S. 2012. Camouflage In Malware: From Encryption To Metamorphism. *International Journal Of Computer Science And Network Security (IJCSNS)*. 12(8):74–83. [Online], Available: http://paper.ijcsns.org/07_book/201208/20120813.pdf.
- Berrueta, E., Morato, D., Magana, E. & Izal, M. 2019. A Survey on Detection Techniques for Cryptographic Ransomware. *IEEE Access*. 7:144925–144944.

- Berrueta, E., Morato, D., Magana, E. & Izal, M. 2020. Open Repository for the Evaluation of Ransomware Detection Tools. *IEEE Access*. 8:65658–65669.
- BlackBerry. 2021. *Cylance*. [Online], Available: <https://www.cylance.com/> [2019, October 05].
- Bromium Labs, Kotov, V. & Rajpal, mantej S. 2014. Understanding Crypto-Ransomware. *Bromium*. 35. [Online], Available: www.bromium.com/sites/.../bromium-report-ransomware.pdf%5CnCached%5CnSimilar%5Cn.
- Dang, D., Di Troia, F. & Stamp, M. 2021. Malware classification using long short-term memory models. *ICISSP 2021 - Proceedings of the 7th International Conference on Information Systems Security and Privacy*. 743–752.
- Dua, S. & Du, X. 2016. *Data Mining and Machine Learning in Cybersecurity*. Auerbach Publications.
- Ernsberger, D., Ikuesan, A.R., Venter, H.S. & Zugenmaier, A. 2017. A Web-Based Mouse Dynamics Visualization Tool for User Attribution in Digital Forensic Readiness. in *9th EAI International Conference on Digital Forensics & Cyber Crime Prague*: Springer Berlin Heidelberg. 1–13.
- Fernando, D.W., Komninos, N. & Chen, T. 2020. A Study on the Evolution of Ransomware Detection Using Machine Learning and Deep Learning Techniques. *IoT*. 1(2):551–604.
- Grégio, A.R.A., De Geus, P.L., Kruegel, C. & Vigna, G. 2013. Tracking memory writes for malware classification and code reuse identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 7591 LNCS:134–143.
- Guarnieri, C. 2014. *Cuckoo Sandbox*. [Online], Available: <https://cuckoosandbox.org/> [2018, October 04].
- Haller, J.T. 2021. *Portable Apps*. [Online], Available: <https://portableapps.com/> [2021, September 02].
- Hampton, N., Baig, Z. & Zeadally, S. 2018. Ransomware behavioural analysis on windows platforms. *Journal of Information Security and Applications*. 40:44–51.
- Knime. 2021. *Knime Analytics Platform*. [Online], Available: <https://www.knime.com/knime-analytics-platform> [2021, June 10].
- Mohlala, M., Ikuesan, A.R. & Venter, H.S. 2018. User attribution based on keystroke dynamics in digital forensic readiness process. in *2017 IEEE Conference on Applications, Information and Network Security, AINS 2017 Vols 2018-Janua*. 1–6.
- Molina, R.M.A., Torabi, S., Sariaeddine, K., Bou-Harb, E., Bouguila, N. & Assi, C. 2021. On Ransomware Family Attribution Using Pre-Attack Paranoia Activities. *IEEE Transactions on Network and Service Management*. 4537(c).
- Newman, L.H. 2018. *Atlanta Spent \$2.6M to Recover From a \$52,000 Ransomware Scare | WIRED*. [Online], Available: <https://www.wired.com/story/atlanta-spent-26m-recover-from-ransomware-scare/> [2019, August 10].
- Riedmiller, M. & Braun, H. 1993. Direct adaptive method for faster backpropagation learning: The RPROP algorithm. *1993 IEEE International Conference on Neural Networks*. 586–591.
- Rughani, P.H. 2017. ARTIFICIAL INTELLIGENCE BASED DIGITAL FORENSICS FRAMEWORK. *International Journal of Advanced Research in Computer Science*. 8(8):10–14.
- SCMagazine. 2019. *Researchers bypass Cylance's AI-based AV solution by masking malware with video game code*. [Online], Available: <https://www.scmagazine.com/home/security-news/researchers-bypass-cylances-ai-based-av-solution-by-masking-malware-with-video-game-code/> [2019, October 05].
- Shah, S.A.R. & Issac, B. 2018. Performance comparison of intrusion detection systems and application of machine learning to Snort system. *Future Generation Computer Systems*. 80(November 2017):157–170.
- Sienko, C. 2021. *InfosecInstitute*. [Online], Available: <https://resources.infosecinstitute.com/category/healthcare-information-security/healthcare-attack-statistics-and-case-studies/ransomware-case-studies-hollywood-presbyterian-and-the-ottawa-hospital/>.
- Singh, J. & Singh, J. 2021. A survey on machine learning-based malware detection in executable files. *Journal of Systems Architecture*. 112(March 2020):101861.
- Singh, A., Ikuesan, A. & Venter, H. 2019a. A context-aware trigger mechanism for ransomware forensics. in *14th International Conference on Cyber Warfare and Security, ICCWS 2019*. 629–638.
- Singh, A., Ikuesan, A.R. & Venter, H.S. 2019b. Digital Forensic Readiness Framework for Ransomware Investigation. *Digital Forensics and Cyber Crime*. 259:91–105.
- Sophos. 2021. *The State of Ransomware 2021*. [Online], Available: <https://secure2.sophos.com/en-us/content/state-of-ransomware.aspx>.
- Subedi, K.P., Budhathoki, D.R. & Dasgupta, D. 2018. Forensic analysis of ransomware families using static and dynamic analysis. *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*. 180–185.
- Vinayakumar, R., Soman, K.P., Velan, K.K.S. & Ganorkar, S. 2017. Evaluating shallow and deep networks for ransomware detection and classification. in *2017 International Conference on Advances in Computing*,

Communications and Informatics, ICACCI 2017 Vols 2017-January. Institute of Electrical and Electronics Engineers Inc. 259–265.

Yuval Nativ, L.L. 2015. *theZoo*. [Online], Available: <https://github.com/ytisf/theZoo> [2018, September 12].

Evaluating the Reliability of Android Userland Memory Forensics

Sneha Sudhakaran¹, Aisha Ali-Gombe², Andrew Case³ and Golden G Richard III¹

¹Department of Computer Science, Louisiana State University, Baton Rouge, USA

²Department of Computer Science, Towson University, Towson, USA

³Board of Directors, Volatility Foundation, Reston, USA

ssudha1@lsu.edu

aaligombe@towson.edu

andrew@dfir.org

golden@cct.lsu.edu

Abstract: Memory Forensics is one of the most important emerging areas in computer forensics. In memory forensics, analysis of userland memory is a technique that analyses per-process runtime data structures and extracts significant evidence for application-specific investigations. In this research, our focus is to examine the critical challenges faced by process memory acquisition that can impact object and data recovery. Particularly, this research work seeks to address the issues of consistency and reliability in userland memory forensics on Android. In real-world investigations, memory acquisition tools record the information when the device is running. In such scenarios, each application's memory content may be in flux due to updates that are in progress, garbage collection activities, changes in process states, etc. In this paper we focus on various runtime activities such as garbage collection and process states and the impact they have on object recovery in userland memory forensics. The outcome of the research objective is to assess the reliability of Android userland memory forensic tools by providing new research directions for efficiently developing a metric study to measure the reliability. We evaluated our research objective by analysing memory dumps acquired from 30 apps in different Process Acquisition Modes. The Process Acquisition Mode (PAM) is the memory dump of a process that is extracted while external runtime factors are triggered. Our research identified an inconsistency in the number of objects recovered from analysing the process memory dumps with runtime factors included. Particularly, the evaluation results revealed differences in the count of objects recovered in different acquisition modes. We utilized Euclidean distance and covariance as the metrics for our study. These two metrics enabled the authors to identify how the change in the number of recovered objects in PAM impact forensic analysis. Our conclusion revealed that runtime factors could on average result in about 20% data loss, thus revealing these factors can have an obvious impact on object recovery.

Keywords: Userland, Memory Dump Acquisition, Reliability, Metric Evaluation

1. Introduction

In a recent survey conducted, the Android operating system makes up 71.9% of the mobile operating system market share (Market Share, 2020). The survey results prove the popularity of the operating system among end-users. Over the last decade, there was a parallel increase in cybercrimes, causing damage that cost up to \$6 trillion in February 2021 (Market Crime, 2021). Digital forensics is a branch of forensic science used for recovering, investigating, and examining digital devices to recover evidence (Auty et al., 2007). Memory forensics is one of the techniques in digital forensics for extracting evidence from digital media that can serve as evidence for solving such cyber-crimes (Sylve et al., 2012). This technique can be an efficient solution to extract evidence and solve cybercrime, thereby making end-users more secure. Among different memory forensics techniques, userland (process memory) memory forensics is one important research area for analyzing applications (app) and activities associated with the app is essential in today's cyber world (Auty et al., 2007). One of the most critical components impacting evidence recovery is process memory acquisition (Pagani et al. 2019). The acquisition must be performed with utmost attention to conquer the challenges that persist. While some research has been done on challenges faced during the memory dump acquisition (Pagani et al. 2019), to the best of our knowledge, there is very little research that focuses on external runtime factors that can affect evidence recovery from process dumps that includes RecOOP (Pridgen et al.). The primary focus of this paper is to conduct a study on the impact of the external runtime factors like the Garbage Collection (GC) and the Process States and finally develop a metric evaluation to assess the reliability of userland memory forensic tools. The external factors like GC during process memory acquisition impact object recovery because the number of objects allocated and recovered without GC occurring is more than objects allocated and recovered after GC's occurrence. The object count difference is primarily because some objects are collected and lost after a GC cycle. The difference in object count was observed in different process states and is described in detail. First, we present a methodology to identify the impact on process memory samples with combinations of external runtime factors on object recovery using some available and free Android userland memory forensic tools (Ali-Gombe et al., 2019) and (Sudhakaran et al., 2020). Next, we evaluated multiple apps and derived a metric

evaluation criterion for measuring the reliability of userland forensic tools on process memory capture acquired by incorporating the runtime factors during the dump acquisition. The study and conclusion deduced from this research were based on the count of objects recovered from process memory dumps with runtime factors included using current Android memory forensic tools. Developing a metric evaluation is needed because of the noticeable changes in the number of objects retrieved. The difference in the number of objects recovered helped us analyze the integrity, consistency, and data loss in different dumps acquired. Therefore, we could better understand userland memory forensic tools' reliability using metrics like Euclidean distance and covariance.

1.1 Contribution

- Examining the impact of Garbage Collection and Process States on evidence recovery in userland memory.
- Quantifying the integrity, consistency, and damage in recovered data.
- Measuring the reliability of userland memory forensics techniques using metrics like Euclidean distance and Covariance.

The rest of the paper is organized as follows: Section 2 presents the Background of this paper; Section 3 provides an overview of our Design and implementation; Section 4 presents the Evaluation of the proposed approach; Section 5 summarizes the Related Literature and finally section 6 presents the Conclusion

2. Background

2.1 Android Runtime Environment

Android executes apps in an application runtime environment called Android Runtime (ART) (Ali-Gombe et al., 2019) (Schwermer et al., 2018). ART introduced significant improvements like GC and better debugging (Schwermer et al., 2018) (Ali-Gombe et al., 2019) (ART Space, 2017). In Android 8, the developers improved ART by introducing features like Concurrent Copying GC (Ali-Gombe et al., 2019) enabling smaller heap sizes and faster object allocation and deallocation (Schwermer et al., 2018). First, the technique uses a concurrent and moving garbage collection algorithm. Utilizing region-based memory allocation (Ali-Gombe et al., 2019), allocated objects are evacuated from a region and subsequently destroyed if and only if the region has live objects whose count is less than some percentage threshold. Also, this algorithm creates a compacting heap by introducing short pauses during collection. It also utilizes a read barrier configuration to ensure mutators never see old versions of objects. This configuration allows threads to efficiently and concurrently access heap objects during collection. The algorithm uses the RegionSpace allocator, and if the use of TLAB is enabled, the system uses the RegionSpaceTlab allocator for movable objects. On newer Android versions, RegionSpaceTlab (Ali-Gombe et al., 2019) is the default for most small object allocations and LargeObjectSpace (Sudhakaran et al., 2020) for large object allocations. Second, the core Android system components and services like ART are built from native code that relies on native libraries written in C/C++ (ART Platform, 2017). Finally, ART's memory management does not provide a memory swap area but instead uses paging mechanisms and file mapping (Soares, A.M.M., de Sousa Jr, RT, 2017). Overall, for end-users, ART is more beneficial than its predecessor Dalvik by offering improved performance and faster application start-up time (ART Dalvik, 2017). In ART, improvements like Foreground and Background collectors are used when an app is in Foreground and Background process states (Ali-Gombe et al., 2019). This research intends to study in detail the GC and process state improvements made in ART and identify how they impact consistency, data loss/integrity, and reliability of forensic evidence recovery tools.

2.2 Object Allocation and Deallocation

In Android, object allocation utilizes memory management algorithms based on the size of the object (AndroidLOS, 2017) (Ali-Gombe et al., 2019) (Sudhakaran et al., 2020). Objects with an allocation size of fewer than 12KB are small objects like primitives, strings, arrays, and other complex objects such as InetAddress and are allocated using the *Alloc()* function (Ali-Gombe et al., 2019). On the other hand, the *AllocLarge()* function allocates large objects above a certain threshold of 12KB (Sudhakaran et al., 2020). The small objects are allocated using the region-based memory management algorithm, and the large objects are allocated using the large object space algorithm. In a region-based algorithm, objects get allocated in specific memory regions. During GC, an entire region is garbage collected if the objects in the corresponding region alive are below a certain threshold. The Large Object Space (LOS) gets allocated in a region in the process memory called Dalvik Large Object Allocation. In ART, LOS uses discontinuous memory mapping, where object allocation regions are not contiguous. LOS allocates objects in the form of arrays of types such as byte, char, string, float, and int

(AndroidLOS, 2017) (Sudhakaran et al., 2020). The objects and the associated references in the allocated spaces are freed when the allocated object is not alive with a *Free()* function (ART Free, 2017).

2.3 Garbage Collection

There are four major GC algorithms designed for ART. The algorithms are *Semi-Space*, *Generational Semi-Space*, *Concurrent Mark Sweep*, and *Concurrent Copying* (Ali-Gombe et al., 2019) (Jones et al., 2016). Beginning with Android 8, the default GC plan is Concurrent Copying (Ali-Gombe et al., 2019). The other GC plan used in ART is Concurrent Mark Sweep (ART Developers, 2017). When utilizing region-based memory allocation, allocated objects in the region having a live object count less than a certain threshold are evacuated from a region and subsequently destroyed (Ali-Gombe et al., 2019). Similarly, allocated objects having an object count greater than the threshold utilize LOS allocation and the objects are collected and destroyed if and only if the object and its reference are no longer alive. In such situations, the function *FinishGC()* (FinishGC, 2017) is enabled indicates that GC has been enabled. Otherwise, the GC is reset or disabled when the *ResetGCPerformanceInfo()* (ResetGC, 2017) function/method is triggered.

2.4 Process State

In an Android system, the process state is the highest-ranking active component within the app that it hosts (Android Process, 2017). Foreground processes have highest priority & empty processes have lowest priority as shown in Figure 1.

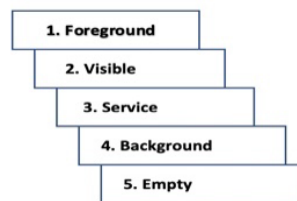


Figure 1: Android Process States

Android's different process states are foreground, background, visible, service, and empty (Android Process, 2017). The different process states are explained below:

2.4.1 Foreground Process

Foreground processes (Android Process, 2017) are the process or the app that is currently an active process running in the Android system and is last to be terminated by the system. A process is in foreground state if it meets one or more of the following conditions given below:

- The process involves activities with user interaction.
- Process hosting a service-connected to user interacting activities.
- Service that is triggered by a function call to *startForeground()*.
- Process that holds services like *onCreate()*, *onResume()* or *onStart()*, *onReceive()* calls.

2.4.2 Visible Process

A process is classified as a 'visible process' if it contains an activity visible to the user while the activity does not involve interaction with the user (Android Process, 2017).

2.4.3 Service Process

Processes that contain a service that has already been started and is currently in execution are classified as service process (Android Process, 2017).

2.4.4 Background Process

These processes contain activities neither visible to the user nor hosting a service. Android maintains a dynamic list of background processes, terminating processes such that processes that were the least recently in the foreground are killed first (Android Process, 2017).

2.4.5 Empty Process

Empty processes no longer contain any active applications but reserve memory space and serve as hosts for newly launched applications (Android Process, 2017).

This research focuses only on Foreground and Background process states.

3. Analysis Setup

3.1 Experimental Setup

In this work, we used the Genymotion Android emulator in the experimental setup as the execution environment (Genymotion, 2016) to evaluate the selected apps. We created Android Virtual Devices (AVD's) for the Samsung S8 emulator running Android 8.0-API 26 and apps chosen from different categories like Browser, Entertainment, SMS, Social media, Vault, Gaming, and malicious apps from Google Playstore(GooglePlay, 2021) and VirusShare(VirusShare, 2021).

All the emulators had 4GB memory, and selected apps were installed and loaded with chat messages, multiple images, text files, videos to simulate a series of actions performed by a user on real devices. We interacted with all the apps selected manually, with a similar sequence of actions conducted on each app to generate consistent activities for evaluation. E.g., The app *com.appstalking.photoeditor* was installed on the Genymotion emulator, and we performed a sequence of actions in both Foreground and Background states. In the Foreground process state, the activities include opening the app and typing a text message, then uploading an image saved in the Genymotion emulator. Next, we uploaded and edited a pdf file, and finally uploaded a video and watched the video. For the background process state, we repeat the same sequence of actions performed on the *com.appstalking.photoeditor* when it was in the foreground, before putting it in the background with a Gmail app made as a foreground app.

3.2 Process Acquisition Modes

As mentioned above, this research leverages two external runtime factors GC and process states. In Figure 2, the memory layout called the vtype (Auty et al., 2007) is explained to understand the different parameters and their memory offsets in the automated script to confirm if the process dump acquired includes the runtime factors or not. The automated script used to check the existence of each runtime factor will be made open source when the paper is published. The heap in the process dump acquired has the vtypes on which this work focuses. The specific vtypes focussed in this work includes GC_collector at location 504; card_table_ at 56; last_gc_type_ at 224; next_gc_type_ at 228; desired_collector_ at 100; block_gc_count_ at 584; block_gc_time_ at 592; gc_plan_ at 440 and concurrent_copying_collector_ at 524 as shown in Figure 2.

```
'Heap' : [0x0388, {
  -
  'gc_collector' : [504, ['']],
  'card_table' : [56, ['']],
  'last_gc_type' : [224, ['']],
  'next_gc_type' : [228, ['']],
  'desired_collector' : [100, ['']],
  'block_gc_count' : [584, ['']],
  'block_gc_time' : [592, ['']],
  'gc_plan' : [440, ['']],
  'concurrent_copying_collector' : [524, ['']],
  -
}],
```

Figure 2: Runtime Structure for External Runtime Factors in Memory

In the case of GC, we identified that when GC is enabled the variables *blocking_gc_count* is 1 and *blocking_gc_time* holds some integer (FinishGC, 2017). While the GC is not enabled, the variables *blocking_gc_count* and *blocking_gc_time* are set to 0 (ResetGC, 2017). In the case of Process State, the variable *desired_collector_* holds a value that indicates if the process was running in the Foreground or Background. The *desired_collector_* is 7 when the app is running in foreground and *desired_collector_* is 8 when the app runs in background. On executing the automated script on all the selected app for analysis each PAM mode gives the corresponding output mentioned below

3.2.1 F-GC

The variable types with values for this memory dump are *desired_collector_* = 0x07; *block_gc_count_* = 1; *block_gc_time_* = <value>.

3.2.2 F-NGC

The variable types with values for this memory dump are *desired_collector_* = 0x07; *block_gc_count_* = 0; *block_gc_time_* = 0.

3.2.3 B-GC

The variable types with values for this memory dump are *desired_collector_* = 0x08; *block_gc_count_* = 1; *block_gc_time_* = <value>.

3.2.4 B-NGC

The variable types with values for this memory dump are *desired_collector_* = 0x08; *block_gc_count_* = 0; *block_gc_time_* = 0.

3.2.5 Clean

This state of process dump acquisition is the fresh state when the app runs with no runtime state triggered. The combination would be a default condition when a user opens the app and interacts in the foreground. This state is acquired every time before acquiring the process memory with the runtime factor triggered. CleanFGC is the process memory dump acquired before acquiring the F-GC combination dump. Similarly, we acquired a clean dump before F-NGC, B-NGC, and B-GC, and the memory acquisition modes were called CleanFNGC, CleanBNGC, and CleanBGC, respectively.

3.3 Process Memory Acquisition For Analysis

Figure 3 depicts the system architecture of this analysis study. The app memory dump is acquired with an automated tool called Memfetch (Memfetch, 2009) with different runtime factors included, and the process dumps acquired are called the process acquisition mode (PAM). The Memfetch tool is more beneficial as this research mainly focuses on userland memory analysis. Memfetch extracts all anonymous memory blocks like the stack and heap used for object allocation (mem pages) and pages used for files and executables mappings (map pages) into distinct binary files called the mem*.bin and the map*.bin respectively (Memfetch, 2009). Memfetch further provides a statistical metadata file that shows the range, size, segment of each acquired memory block in a file called the mfetch.lst. These files are then analyzed by forensic tools to recover all the objects. In our research, we focus specifically on two runtime factors, the GC and Process States.

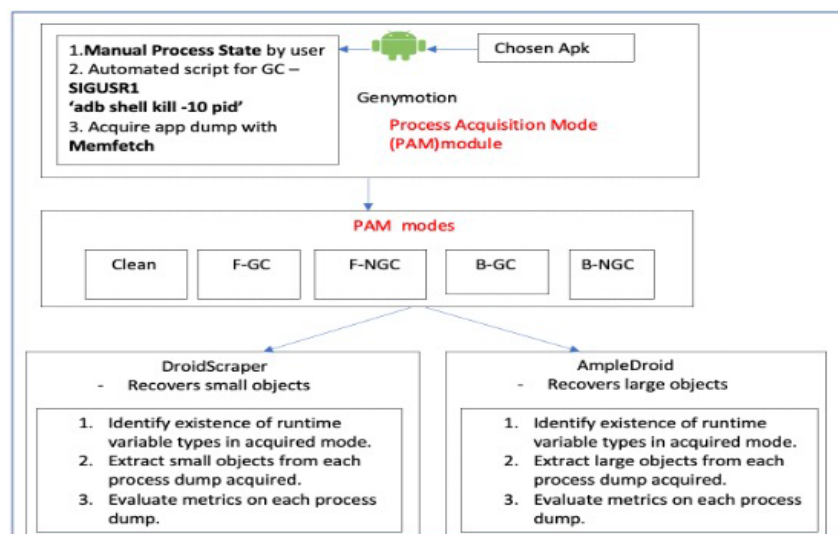


Figure 3: Design Architecture for Evaluation

In the automated script (<https://github.com/ssudha1/metrics/tree/main>), before we used Memfetch for acquiring process dumps, we initiated the two runtime factors. The GC was forced from the Android shell using 'adb shell kill -10 PID' where 'kill -10' signals SIGUSR1 (ForceGC, 2013), and PID is the process id of the application. However, the process states during memory acquisition were based on the manual setting by the end-user. We focus on the Foreground and Background process states which predominantly relate to user input. These states occur when the user acquires the process dump by running the automated script. The Foreground state is where the user interacts with the application, and the Foreground Collector gets triggered. The Background process state is where the application under execution is not visible to the user, and the Background Collector is triggered. E.g When the user starts the app in the Foreground process state and then executes the automated script (<https://github.com/ssudha1/metrics/tree/main>), it generates a process dump without GC enabled called

Foreground and No GC enabled (F-NGC) followed by a dump with GC enabled called Foreground and GC triggered (F-GC). Similarly, when the user starts the app in the Background state and the automated script (<https://github.com/ssudha1/metrics/tree/main>) is executed, the process dump acquired would be initially the dump without GC enabled called Background and No GC enabled (B-NGC) followed by a dump with GC called as the Background and GC enabled (B-GC).

3.4 Analysis Tools

3.4.1 DroidScraper

Ali-Gombe et al. proposed DroidScraper (Ali-Gombe et al., 2019), a tool for analyzing the ART RegionSpace memory allocation to extract small objects from process dumps. This tool recovers runtime data

structures by focusing on the recovery of small objects allocated in the RegionSpace. DroidScraper is a userland in-memory object recovery and reconstruction tool to extract runtime artifacts from Android process memory space. The evaluation of DroidScraper has shown that it can recover in-memory data allocated using Android's RegionSpace by using RegionSpaceTlab allocator for recovering small objects with a recovery percentage of almost 90%. Also, DroidScraper can reconstruct and recover objects, thereby detecting evidence of file and network activities, database accesses, and recovery of cryptographic keys.

3.4.2 AmpleDroid

Sudhakaran et al. proposed a tool AmpleDroid (Sudhakaran et al., 2020) a tool to analyze an Android app dump to extract the large objects allocated in the specific memory region called LargeObjectSpace. The LargeObjectSpace allocation is studied to identify how the objects above a certain threshold are stored in userland memory. AmpleDroid performs a complete process memory analysis on Android version 8 large object memory allocation and extracts multimedia and text files that other tools cannot currently retrieve. This tool is used in forensic investigations to provide an overall idea of how large object files (text, video, image, etc.) are allocated in an app's memory and a high recovery percentage of approximately 91%.

Lastly, in this experimental setup the forensic tools - DroidScraper (Ali-Gombe et al., 2019) and AmpleDroid (Sudhakaran et al., 2020) are used to recover the objects allocated in each acquisition mode. The final objects extracted from all the dumps will include a set of objects extracted from an Android app memory. Finally, the authors were able to identify the objects recovered in each dump and determine how reliable these forensic tools are in recovering objects from process memory.

4. Evaluation

We evaluated 30 apps in different process acquisition modes (cleanFGC, F-GC, cleanFNGC, F-NGC, cleanBGC, B-GC, cleanBNGC, B-NGC), and the results are shown in Figure 4. We evaluated 15 benign apps from different app categories: Browser, Editing, Entertainment, Gaming, SMS, Social Media and 15 malicious apps. Upon analysis, we identified that the process dumps with the acquisition modes - no GC and both process states (Foreground and background) extracted more small and large objects than the process dumps with acquisition modes GC and both process states. We used two metrics, Euclidean distance and covariance, to evaluate the data loss. The Euclidean distance calculated here gives the percentage of dissimilarity between two acquisition modes, e.g. (Data recovered in a clean state, data retrieved with runtime factor triggered). We used the covariance metric to examine the variability of data retrieved in different process acquisition modes, e.g., the variability of data recovered in F-NGC, F-GC, B-NGC, and B-GC. The results in Figure 4 prove a greater probability for objects and their references get collected in the case of process acquisition modes F-GC and B-GC. On analysis with DroidScraper and AmpleDroid, we identified that the number of recovered objects in modes F-GC and B-GC was lower than F-NGC and B-NGC. In Figure 4 shows a significant difference in the small and large objects from all acquisition modes in certain benign entertainment apps like *com.vid007.videobuddy*, *com.redbox.tv* and *com.ezscreenrecorder*. The range of data loss calculated using Euclidean distance is approximately 10-20% in the case of GC acquisition modes while 0-10% in the case of No-GC acquisition modes. The data loss in terms of covariance ranges between .06 to .09 in (6-9% varying) both small and large objects. Therefore, we identified the change to be consistent in the case of entertainment apps. In the case of SMS apps, we identified the Euclidean distance as 7-17% and 0-7% for GC and No-GC process acquisition modes, respectively. We also determined that the covariance value was between .04 to .09 (4-9% varying). However, this consistency was not observed in Browser and Editing apps as we observed a Euclidean metric ranging from 0-19% in the case of all the acquisition modes and covariance variability from 0 to 11 (0 to 11% varying). Therefore, we could not identify

a consistent range of percentage data loss but can conclude that the loss of data is restricted to 20% in the case of all benign and malicious applications. Thus, the garbage collected objects removed from memory might or might not serve as a loss of a critical piece of evidence during object recovery. In any case, if such pieces of evidence are not recovered, it eventually leads to issues that may question the reliability of forensic investigation tools. Such an object recovery can miss out on evidence that can also cause innocent users to be convicted. Therefore, on analyzing the difference in the count of objects recovered it is clear from in Figure 4 that the objects recovered using B-GC acquisition mode were fewer and had greater Euclidean distance compared to other modes. This also made us understand that among the process states lesser objects were recovered from Background compared to objects recovered from dumps in foreground. For E.g, on comparing F-GC and B-GC, more objects were recovered in F-GC compared to B-GC.

Packagename	App Category	Large Clean-Fgc	Large F-NGC	Large Euclid%	Large Clean-Fgc	Large F-GC	Large Euclid%	Large Clean-Bgc	Large B-NGC	Large Euclid%	Large Clean-Bgc	Large B-GC	Large Euclid%	Covariance
com.brave.browser	Browser	18	18	0.00%	19	16	15.79%	17	17	0.00%	17	14	17.65%	0.11
com.aniplex.fategrandorder	Browser	16	16	0.00%	16	14	12.50%	15	15	0.00%	15	13	13.33%	0.09
org.mozilla.firefox	Browser	12	12	0.00%	12	11	8.33%	12	12	0.00%	11	11	8.00%	0.05
com.appetizing.photoeditor	Editing	18	18	0.00%	17	16	5.88%	17	17	0.00%	17	15	11.76%	0.08
de.vsmmedia.imagesize	Editing	16	16	0.00%	16	14	12.50%	16	15	6.25%	16	13	18.75%	0.09
com.videobuddy	Editing	11	11	0.00%	11	11	0.00%	11	11	0.00%	11	11	8.00%	0
com.redbox.tv	Entertainment	10	10	0.00%	10	9	10.00%	10	10	0.00%	10	9	18.00%	0.06
com.screensrecorder	Entertainment	28	28	0.00%	28	27	3.57%	28	28	0.00%	28	24	14.29%	0.07
com.mp3.editor.music.cut.ringtones	Entertainment	30	29	3.33%	30	27	10.00%	30	28	6.67%	30	25	16.67%	0.06
com.tencent.ligite	Gaming	14	14	0.00%	13	11	15.38%	13	12	7.69%	13	11	15.38%	0.12
com.anonymoustexting	Sms	14	14	0.00%	14	13	7.14%	14	14	0.00%	14	12	14.29%	0.07
com.sms.plus.app	Sms	18	17	5.56%	18	15	16.67%	17	16	5.88%	17	15	11.76%	0.06
co.kittech.messenger	Sms	34	33	2.94%	34	31	8.82%	34	33	2.94%	34	31	8.82%	0.04
com.nexo.chatmessenger.ui	Social Media	29	29	0.00%	29	28	3.45%	29	29	0.00%	29	27	6.90%	0.03
gallery.hidepictures.photovault.lockgallery	Vault	19	19	0.00%	19	17	10.53%	19	18	5.26%	19	17	18.53%	0.05
com.nemo.videomate	Malware	54	54	0.00%	54	48	11.11%	54	51	5.56%	50	46	8.00%	0.07
com.gihoo.appstore	Malware	37	37	0.00%	37	31	16.22%	35	33	5.71%	35	30	14.29%	0.09
com.wBestFreeGameCheatCodesGlitches	Malware	9	9	0.00%	9	8	11.11%	9	9	0.00%	9	8	11.11%	0.07
com.ss.android.article.news	Malware	12	12	0.00%	11	9	18.18%	11	10	9.09%	11	9	18.18%	0.14
com.namad.instagram	Malware	8	8	0.00%	8	8	0.00%	8	8	0.00%	8	8	8.00%	0
com.android.tencent.zdevs.bah	Malware	12	12	0.00%	12	11	8.33%	11	11	0.00%	11	10	9.09%	0.07
ru.delivery.collapse	Malware	22	22	0.00%	22	20	9.09%	21	20	4.76%	21	19	9.52%	0.06
com.isyiv.kxbimwc.r	Malware	9	9	0.00%	9	8	11.11%	9	9	0.00%	9	8	11.11%	0.07
org.doviz.cevir	Malware	15	15	0.00%	15	13	13.33%	14	14	0.00%	14	12	14.29%	0.1
operator.italia	Malware	13	13	0.00%	13	11	15.38%	13	12	7.69%	13	11	15.38%	0.08
com.marjansb1.thanksgiving	Malware	17	17	0.00%	17	15	11.76%	17	16	5.88%	17	14	17.65%	0.08
com.amazon.mShop.android.shopping.ha	Malware	18	17	5.56%	17	15	11.76%	17	16	5.88%	17	14	17.65%	0.08
com.androiddoctor.battery	Malware	10	10	0.00%	10	9	10.00%	10	9	10.00%	10	9	18.00%	0.05
com.maxauto.maxipusap	Malware	8	8	0.00%	8	7	12.50%	8	8	0.00%	8	7	12.50%	0.08
com.plato.dovizim	Malware	26	26	0.00%	26	24	7.69%	26	25	3.85%	26	22	15.38%	0.07
Packagename	App Category	Small Clean-Fgc	Small F-NGC	Small Euclid%	Small Clean-Fgc	Small F-GC	Small Euclid%	Small Clean-Bgc	Small B-NGC	Small Euclid%	Small Clean-Bgc	Small B-GC	Small Euclid%	Covariance
com.brave.browser	Browser	39159	39002	0.40%	38162	33241	12.50%	38831	37198	4.21%	38021	33108	12.92%	0.08
com.aniplex.fategrandorder	Browser	41629	40998	1.52%	40864	35912	12.12%	40799	40001	1.98%	40006	35482	11.31%	0.07
org.mozilla.firefox	Browser	35197	35098	0.28%	35481	30200	14.88%	33217	30777	7.35%	32183	28939	16.24%	0.11
com.appetizing.photoeditor	Editing	30189	30027	0.60%	30819	26108	14.73%	30201	29009	3.95%	30765	27885	9.36%	0.06
de.vsmmedia.imagesize	Editing	48910	48008	1.84%	48529	41198	15.11%	48873	42107	13.84%	48003	41298	13.97%	0.08
com.videobuddy	Editing	37803	36998	1.61%	37052	30008	19.01%	36458	35002	3.99%	36897	32542	11.88%	0.09
com.redbox.tv	Entertainment	17888	17092	4.46%	17167	13998	18.46%	17165	16917	1.44%	17165	15078	12.17%	0.09
com.screensrecorder	Entertainment	41332	40975	0.86%	41134	35882	12.77%	41221	39887	3.24%	41225	35887	13.48%	0.07
com.mp3.editor.music.cut.ringtones	Entertainment	28967	28103	2.98%	28886	23881	17.33%	29040	27669	4.72%	28887	24997	13.41%	0.08
com.tencent.ligite	Gaming	21761	21668	0.43%	21322	17682	17.07%	21325	20534	3.71%	21583	17992	16.64%	0.1
com.anonymoustexting	Sms	38910	38885	0.12%	38654	33265	13.94%	38912	37798	2.86%	38912	34004	12.61%	0.08
com.sms.plus.app	Sms	31876	31789	0.34%	31807	26573	16.46%	31885	29880	6.29%	31885	28358	17.34%	0.09
co.kittech.messenger	Sms	53844	53281	1.05%	52967	47562	10.20%	53083	50832	4.24%	53001	47987	18.06%	0.06
com.nexo.chatmessenger.ui	Social Media	43997	42681	3.04%	42960	38038	11.50%	43007	39005	9.31%	43603	36431	16.45%	0.07
gallery.hidepictures.photovault.lockgallery	Vault	39870	38974	2.25%	38994	32198	17.43%	39007	37420	4.07%	39105	35996	7.95%	0.08
com.nemo.videomate	Malware	43556	43183	0.86%	43199	38241	11.46%	43277	42176	2.54%	43004	37881	11.91%	0.07
com.gihoo.appstore	Malware	44472	44389	0.18%	44321	38626	12.17%	44188	43227	2.17%	44183	39881	9.74%	0.06
com.wBestFreeGameCheatCodesGlitches	Malware	27881	26990	4.27%	26880	23992	11.07%	26997	24081	10.80%	27034	22780	15.81%	0.07
com.ss.android.article.news	Malware	55135	55121	0.03%	55002	50165	8.79%	49388	48131	2.55%	49388	42752	13.44%	0.1
com.namad.instagram	Malware	28762	28006	2.63%	28661	23769	17.07%	29901	29548	1.18%	28995	23996	17.24%	0.11
com.android.tencent.zdevs.bah	Malware	33287	33186	0.27%	33192	27379	17.51%	33267	32669	1.80%	33215	27549	17.06%	0.1
ru.delivery.collapse	Malware	41880	40998	2.11%	41021	35201	14.19%	42005	38109	9.28%	41785	36626	12.34%	0.07
com.isyiv.kxbimwc.r	Malware	25990	25079	3.51%	25321	20438	19.28%	24468	22567	7.77%	25083	20270	19.19%	0.1
org.doviz.cevir	Malware	43871	42889	2.24%	42765	37719	11.80%	41328	39874	3.52%	40696	35489	13.43%	0.08
operator.italia	Malware	41876	41006	2.08%	41058	35791	12.83%	41078	38452	6.39%	41198	36083	12.42%	0.06
com.marjansb1.thanksgiving	Malware	28894	28549	1.19%	28651	23956	16.39%	28709	28003	2.86%	28771	23993	16.61%	0.1
com.amazon.mShop.android.shopping.ha	Malware	29983	29156	2.76%	29318	24147	17.64%	29845	27452	8.02%	29845	24176	18.99%	0.09
com.androiddoctor.battery	Malware	19860	19778	0.41%	19862	16281	17.20%	19837	18995	3.27%	19732	16032	18.75%	0.11
com.maxauto.maxipusap	Malware	19899	19118	3.92%	19650	16592	15.56%	19775	18043	8.76%	19532	16553	15.25%	0.07
com.plato.dovizim	Malware	45668	44774	1.96%	45801	39667	12.69%	44997	43219	3.95%	45097	39842	11.65%	0.06

Figure 4: Evaluation Result

4.1 Recovery Consistency and Data Loss

On analysis, we identified that the results in Figure 5 reveal that we retrieved more non-Gui objects (Class name, text, etc.) in case of benign (apps selected from each category) and malicious applications than Gui and network objects. When we compared the count of Gui objects recovered from dumps with GC triggered and dumps with No-GC triggered, we identified that in the Foreground process dumps, the Editing, Entertainment, and SMS apps had an average of 19% reduction while the malware apps had an average of 24% reduction. When we compared the Gui object count for B-NGC and B-GC, in the case of Editing, Entertainment and SMS, and Malware apps, the reduction in object recovered was 19% !9.5%, 24.5%, and 21%, respectively. When we compared the count of non-Gui objects recovered from dumps with GC triggered and dumps with No-GC triggered, we identified that in the Foreground process dumps, the Editing, Entertainment, and SMS apps had an average of 20% reduction. In comparison, the malware apps had an average of 22% reduction. When we compared the Gui object count for B-NGC and B-GC, in the case of Editing, Entertainment and SMS, and Malware apps, the reduction in object recovery count was 20%.

Similarly, the reduction in Network objects recovered ranges between 18% to 23% in all the apps studied in this research. The histograms in Figure 6 include component-based recovery from a few apps selected from each app category. Figure 6 depicts that we retrieved more Android Services than Android activity components and resource files on plotting apps from each category like editing, malware, SMS, entertainment.

Acquisition Mode	Package Name	Package type	Large	Gui	Non-Gui	Network
F-NGC	com.appstalking.photoeditor	Editing	18	1728	24721	205
	de.vsmidia.imagesize	Editing	16	2552	38119	429
	com.redbox.tv	Entertainment	10	1273	11883	128
	com.ezscreenrecorder	Entertainment	28	2181	30449	398
	com.anonymoustexting	SMS	14	2003	28771	361
	com.smsplus.app	SMS	17	1549	21553	318
	com.androiddoctor.battery	Malware	10	1161	10001	102
	com.maxauto.maxiplusap	Malware	8	1004	9883	111
Acquisition Mode	Package Name	Package type	Large	Gui	Non-Gui	Network
F-GC	com.appstalking.photoeditor	Editing	16	1400	15281	134
	de.vsmidia.imagesize	Editing	14	1994	31108	317
	com.redbox.tv	Entertainment	9	991	8329	99
	com.ezscreenrecorder	Entertainment	27	1772	25771	253
	com.anonymoustexting	SMS	14	1558	23165	321
	com.smsplus.app	SMS	13	1287	16459	276
	com.androiddoctor.battery	Malware	9	792	7648	81
	com.maxauto.maxiplusap	Malware	7	839	7805	87
Acquisition Mode	Package Name	Package type	Large	Gui	Non-Gui	Network
B-NGC	com.appstalking.photoeditor	Editing	17	1806	21904	191
	de.vsmidia.imagesize	Editing	15	2469	36881	431
	com.redbox.tv	Entertainment	10	1309	12115	133
	com.ezscreenrecorder	Entertainment	28	2471	31985	405
	com.anonymoustexting	SMS	14	2096	29001	380
	com.smsplus.app	SMS	16	1602	20998	338
	com.androiddoctor.battery	Malware	9	1201	11998	119
	com.maxauto.maxiplusap	Malware	8	1188	12006	97
Acquisition Mode	Package Name	Package type	Large	Gui	Non-Gui	Network
B-GC	com.appstalking.photoeditor	Editing	15	1489	16098	148
	de.vsmidia.imagesize	Editing	13	1992	31116	338
	com.redbox.tv	Entertainment	9	997	9003	102
	com.ezscreenrecorder	Entertainment	24	2097	25993	297
	com.anonymoustexting	SMS	12	1496	25669	319
	com.smsplus.app	SMS	15	1274	16889	269
	com.androiddoctor.battery	Malware	9	992	8003	97
	com.maxauto.maxiplusap	Malware	7	907	8251	83

Figure 5: Component based Analysis

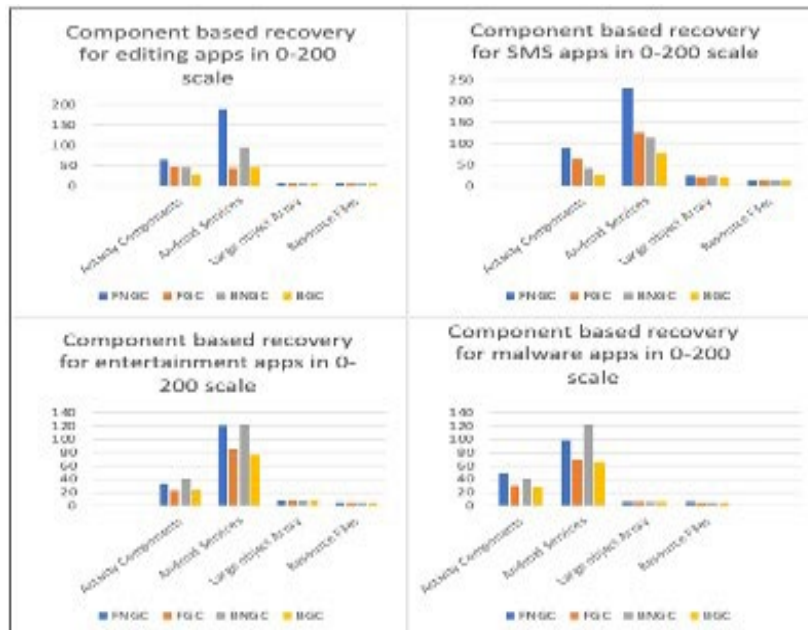


Figure 6: Histogram analysis of various files in selected apps

5. Related Literature

The analysis of software tools' reliability is inevitable in today's digital world (Ivanov, et. al, 2018)}. The reliability assessment is more critical if the analyzed tool is used for solving cyber-related cases. Reliability assessment of such tools is required to analyze if the tool is unreliable, then the evidence extracted by the tool also becomes unreliable, thereby ending up generating incorrect evidence. Hence the reliability assessment of such tools is mandatory, and this research provides a methodology for one such assessment. We identified multiple published works on reliability assessment; however, only a few were related to reliability testing of memory dump acquisition for forensic analysis. Most of the reliability measurement works were based on multiple approaches to develop software reliability growth models using techniques like fuzzy models, regression analysis, neural networks, and machine learning (Ivanov et al., 2018). Another technique for reliability testing is by methodologies like training and testing neural networks (Fisch et. al, 2010) (Park et. al, 2013). However, in this work, we focus on issues about software that focus on the extraction of all types of objects (small and large) from smartphone app memory. This is difficult due to the challenges faced during process dump acquisition (Sylve et. al, 2012) (Pagani et. al, 2019) (Schatz et. al, 2007). There are many works associated with the recovery of forensic data. However, only a few works address the challenges faced by forensic investigators with memory acquisition (Pagani et. al, 2019), are more useful in this work. Among different memory dump acquisition techniques mentioned in works like Volatility (Auty et. al, 2007), (Schatz et al., 2007) - Schatz, in his work, explained his technique for reliable volatile memory dump recovery. Also, Pagani et al., in their work, explained how memory forensics should consider the time in which each memory dump was acquired (Pagani et. al, 2019). Pagani et al. in their work provided a way to assess the reliability of a result obtained thereby minimizing the effect of the acquisition time or detect inconsistencies in the data. While our research closely relates to the idea of reliability (Huelsenbergen et. al, 1993) but focuses primarily on external runtime factors, we could not find any work that mentioned runtime factors as critical during memory dump acquisition. Therefore, our research provides a new research dimension of focusing on the critical memory dump acquisition. Hasanbadi et al. proposed an approach to determine approximately how much sequential memory acquisition at a designated time-intervals can mitigate the current challenges in memory forensics (Hasanbadi et al., 2018).

6. Conclusion

In this paper, the methodology identified two runtime factors that can impact Android application dump acquisitions. We presented the analysis results highlighting the changes in the count of objects recovered in every app dump analyzed using userland memory forensic tools. We evaluated multiple apps and identified a difference in object recovery rate during the analysis of app dumps with the runtime factors included during memory acquisition. Finally, we calculated the reliability of userland memory forensic tools using Euclidean distance and covariance metrics. Our evaluation of 30 apps (benign and malicious apps) shows these process

states can impose data loss of approximately 20% with a metric Euclidean distance and less than 18% data loss with covariance metric. Furthermore, our comparative analysis found that the count of objects recovered from Foreground acquisitions modes are greater than objects recovered from Background acquisition modes in most of the apps analyzed. The userland forensic analysis's reliability study was conducted on both small and large object recovery tools. The result will be more reliable in object recovery for a forensic investigator who can extract all the objects allocated in memory from the app startup. Also, this research highlights runtime factors and helps investigators explain the environment during memory acquisition, thereby providing a better understanding of reliability factors when dealing with forensic tools during memory dumps acquisition.

References

- AndroidLOS, 2017 http://androidxref.com/8.0.0_r4/xref/art/runtime/gc/space/large_object_space.cc#180 [Online: accessed 13-March 2021]
- Market Crime, 2021 "300 terrifying cybercrime and cybersecurity statistics trends" Online: <https://www.comparitech.com/vpn/cybersecurity/protect/discretionary>
- Android 8.0 ART Improvements, 2017 Online: <https://source.android.com/devices/tech/dalvik/improvements>. [Online: accessed 13-March 2021].
- Android platform architecture, 2017 URL: <https://developer.android.com/guide/platform>. [Online: accessed 10-March 2021].
- Androidxrefspace URL: http://androidxref.com/8.0.0_r4/xref/art/runtime/gc/space/. [Online: accessed 10-March 2021].
- AndroidLOS, 2017 http://androidxref.com/8.0.0_r4/xref/art/runtime/gc/space/large_object_space.h
- Android Dal, 2017 URL: <https://source.android.com/devices/tech/dalvik/gc-debug>.
- Zalewski, M. (2002) from <http://lcamtuf.coredump.cx/soft/memfetch.tgz> [Online; accessed 17-February 2021].
- Market Share, 2020 <https://gs.statcounter.com/os-market-share/mobile/worldwide>. [Online: accessed 10-March 2021].
- Android Process, 2017, URL: <https://learnswithandroid.blogspot.com/2017/12/android-process-states.html> [Online: accessed 10-March 2021].
- Ali-Gombe, A., Sudhakaran, S., Case, A., Richard III, G.G., 2019. "DroidScraper: A tool for android in-memory object recovery and reconstruction", in: 22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID) 2019), pp. 547–559.
- Auty, M., Case, A., Cohen, M., Dolan-Gavitt, B., Ligh, M.H., Levy, J., Walters, A., (2007). "Volatility-an advanced memory forensics framework".
- Fisch, D., Hofmann, A., Sick, B., 2010. "On the versatility of radial basis function neural networks: A case study in the field of intrusion detection", Information Sciences 180, 2421–2439.
- Hasanabadi, S.S., Lashkari, A.H. and Ghorbani, A.A., 2018, October. "The Next Generation of Robust Linux Memory Acquisition Technique via Sequential Memory Dumps at Designated Time Intervals." In *2018 International Carnahan Conference on Security Technology (ICCST)* (pp. 1-6). IEEE.
- ResetGC, 2017, http://androidxref.com/8.0.0_r4/xref/art/runtime/gc/heap.cc#1147
- Ivanov, V., Reznik, A., Succi, G., 2018, "Comparing the reliability of software systems: A case study on mobile operating systems". Information Sciences 423.
- Jones, R., Hosking, A., Moss, E., 2016, "The garbage collection handbook: the art of automatic memory management", CRC Press.
- Pagani, F., Fedorov, O., Balzarotti, D., 2019, "Introducing the temporal dimension to memory forensics. ACM Transactions on Privacy and Security (TOPS) 22, 1–21.
- Park, B.J., Oh, S.K., Pedrycz, W., 2013, "The design of polynomial function-based neural network predictors for detection of software defects", Information Sciences 229, 40–57.
- Schatz, B., 2007, "Toward reliable volatile memory acquisition by software" URL: <https://www.dfrws.org/2007/proceedings/p126-schatz.pdf>. [Online: accessed 10-March 2021].
- Schwermer, P., 2018, "Performance evaluation of Kotlin and Java on Android Runtime"
- Soares, A.M.M., de Sousa Jr, R.T., 2017. "A technique for extraction and analysis of application heap objects within android runtime (ART)", in: ICISSP, pp. 147–156.
- Sudhakaran, S., Ali-Gombe, A., Orgah, A., Case, A. and Richard, G.G., 2020, December. "AmpleDroid recovering large object files from Android application memory". In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE.
- Sylve, J., Case, A., Marziale, L., Richard, G.G., 2012. "Acquisition and analysis of volatile memory from android devices". Digital Investigation 8, 175–184
- Pridgen, A., Garfinkel, S., Wallach, D. S., "Picking up the trash: Exploiting generational GC for memory analysis", DFRWS, 2017
- Memfetch, 2009, <http://shellcoders.blogspot.com/2009/05/using-memfetch-page-37.html>
- Genymotion Emulator, 2016, <https://www.genymotion.com> [Online: accessed 10- September 2021]
- FinishGC, 2017, http://androidxref.com/8.0.0_r4/xref/art/runtime/gc/heap.cc#2804
- GooglePlay, 2021, https://play.google.com/store?hl=en_US&gl=US
- VirusShare, 2021, <https://virusshare.com>
- ForceGC, 2013, <http://www.codeflow.fi/2013/09/21/force-gc-from-shell/>

The Cumulative Cyber Deterrence

Maija Turunen and Martti J. Kari

Finnish National Defence University, University of Jyväskylä, Finland

maijaturunen@yahoo.fi

martti.j.kari@jyu.fi

Abstract: The cumulative cyber deterrence can be seen as a concept in which increasing the weight of different means and their use increases the deterrent effect on a common level or on selected adversaries. Cumulative cyber deterrence may include all traditional options of deterrence, and can be active or passive. Active deterrence can be characterized as targeting specific threats and actors, as a deterrent consisting of several different methods, while passive deterrence is a form of deterrence commonly targeted at all the potential adversaries. The cumulative cyber deterrence can be an independent type of deterrence or part of a state's overall deterrence. This paper approaches the concept of cumulative cyber deterrence from a military perspective. The purpose is to determine what factors can be formed by cumulative cyber deterrence. It describes how cumulative deterrence will change and be affected and what problems can be associated with that concept. The aim is to find answers to these questions by looking at the way how Israel and Russia use cumulative cyber deterrence as part of their overall deterrence. In its theoretical context, this paper is based on the theory of the character of war. Through the theory of character of war and utilizing the concept of reflexive control, an attempt is made to explain the position of cumulative cyber deterrence as part of overall deterrence. Integrative literature analysis has been used as the research method. The key conclusion of the paper is that creating a credible cyber deterrent is an affect and cost-effective way to increase overall deterrence. However, this presupposes that the state also has offensive cyber methods at its disposal and is able to credibly communicate their existence and the will to use them if necessary. The concept of cumulative cyber deterrence depends on the other means of deterrence available to the state. Both Israel and Russia have all these qualities. A key difference in the deterrence strategies of these states is that Israel uses cumulative methods to make it clear where the red lines are, while Russia's strategic goal is to blur them.

Keywords: Cyber Deterrence, Russia, Israel, Red Lines

1. Introduction

Cyberspace is a dynamic, aggregated and rapidly changing battle domain with a wide range of actors and threats. The cyber deterrent may be part of the state's overall deterrence or just part of military deterrence. The cyber deterrent may be targeted solely against attacks in the cyber domain, or it may include a punitive element against hostile acts in other domains as well.

An overall deterrence or restraint created by the state can consist of a combination of political, military, informational, diplomatic, financial, intelligence, economic, legal, and developmental methods and measures. The use of cyber measures is generally considered a military, intelligence or information method, even though the objectives may be, for example, political or diplomatic. In this paper, cyber methods are discussed as part of overall military deterrence. The hypothetical question is if cyber methods can create an detached deterrent that is independent from other forces or weapon systems, for example, if an adversary is led to believe that someone has access to its C2ISR (Command, Control, Intelligence, Surveillance, Reconnaissance) systems and hostile behavior will result in punishment in those systems.

In the cyber warfare, the targets of operations can be varied: attacking the information available to the adversary, blurring the situational awareness and perception of the war character; interfering with or paralyzing the operation of the adversary's command, control, surveillance and weapon systems; obstructing or impeding the use of the adversary's critical infrastructure; (Arquilla & Ronfeld, 1993, 31, 33) creating an anti-access/area denial (A2/AD) battle domain and/or creating access to an adversary's A2/AD mode, and so on. In the concept of cumulative cyber deterrence, these targets can be valued and the force used for countermeasures adjusted accordingly.

The authors consider that cyber deterrence can be seen as consisting of the strategic, operational, tactical and technological capabilities of the state or its allies, the ability to develop them, the credible communication of their existence and the willingness to use them when necessary. These capabilities, the principles of use them, and targets vary at different levels. The state can strategically communicate its cyber deterrence through, for example, statements from its political or military leadership (will), strategy papers, legislation (plans), military exercises or even its operations (capability).

When acting at the operational level, the deterrent can be controlled using defensive, offensive or command and control system operations or, more specifically and especially in connection with deterrence, flexible defence operations (FDO) and flexible response operations (FRO). They both provide the ability to scale up (escalate) or de-escalate based on continuous assessment of an adversary's actions and reaction (JP 3-0, 2017, VIII-9). U.S. Department of Defense Science Board (2017, 15) highlights the importance to maintain scalable offensive cyber capabilities as an integral part of its cyber deterrence posture. This requires the credibility and explainability of cyber counterattacks and responses. On the operational level, the timing and operational complexity of cyber operations are emphasized (Schulze, 2020, 193).

The tactical level capabilities like operations control- and execution are typically unique and the use of the force and cyber methods by situation based. The problem with using tactical level capabilities is that they can usually only be used once (Schulze, 2020, 192). The existence, or lack thereof, of technical capabilities is well suited to the clandestine and covert world of cyber influencing. One way to get information about technological capabilities is analyzing completed operations or something about the potential capabilities of a state can be deduced by analyzing its cooperative and alliance relations.

1.1 Theoretical background and research methodology

This research applies the theory of the character of war to cumulative cyber deterrence. The character of war can be defined as the perceptions in the international system of the nature, needs and possibilities of the use of armed forces, as well as the effective principles and operating models of the armed forces (Raitasalo et al, 2008, 9). The use of the concept of reflexive control seeks to influence the adversary's situational awareness and, more broadly, the perception of the character of war. The concept of reflexive control has been designed to be executed at strategic, operational, and tactical levels (Kasapoglu, 2015, 5).

The concept of reflexive control consists of analyzing an adversary's behavior, ideas and practices and then trying to influence them. Reflexivity refers to the creation of a particular pattern of behavior in a system that is sought to be influenced and controlled (the objective system). It takes into account the fact that the object system has a situation picture that is also assumed to affect the adversary's leadership or sociopolitical system. Reflexive control utilizes moral, psychological, and other factors as well as the personal qualities of leaders. The concept of reflexive control can be seen as creating a framework for action aimed at influencing the action of an adversary so that it voluntarily begins to act in the desired way (Thomas, 2004, 237–242).

The concept of reflexive control plays a central role in Russian art of warfare, which is also influenced by strategic culture. Kari (2019, 71; Johnston, 1995) explains strategic culture as a set of persistent and consistent historical patterns of how state leadership thinks about the use of force to achieve political goals. The preferences originate in the historical experiences related to the threat and use of force by the state and are influenced by the philosophical, political, cultural, and cognitive experiences and characteristics of the state.

Integrative literature analysis has been used as the research method. By integrating and analyzing the literature on deterrence in general and cyber deterrence in particular around the research question "Cumulative Cyber Deterrence", the aim is to position the examination of the research question as part of the scientific debate on the topic.

2. The Concept of Cyber Deterrence

The concept of cyber deterrence has been discussed in the literature: can it be an independent deterrent, part of another deterrent or does it exist at all? According to Soesanto and Smeets (2020, 392-394), views on this issue can be roughly divided into three groups: (1) cyber deterrence functions akin to conventional deterrence; (2) cyber deterrence features as unique issues because cyberspace is markedly different from the traditional domains; and (3) cyber deterrence is impossible.

Analyzing these views, the authors estimated that the cyber deterrent in its objectives corresponds to a conventional deterrent. Generally, deterrence means preventing an adversary from taking undesired action. The general theory of deterrence is defined as the use of means of decisive influence over an adversary's decision-making. Traditional deterrence is based on an adversary's perception that a threat of retaliation exists, or the planned attack or other action cannot be successful or the costs of the attack outweigh the benefits (Jasper, 2018, 161). Furthermore, Goodman (2010, 108) highlights that "in addition to strong denial measures, classical

deterrence theory demands that penalty measures be certain, severe, and immediate; however, cyber deterrence emphasizes certainty more so than severity or immediacy.”

Compared to land, air, sea, and space domains, the cyber domain is different, because changes there can occur very fast and comes unnoticed. There are more potential adversaries and the cyber weapons they use may be unknown. It is precisely because of this constant activity in cyberspace and its rapid pace of change, for example, that Williams (2017) is skeptical of the deterrent’s functionality in cyberspace as an operational environment because he sees an absence of action as an indication of deterrent effectiveness. Instead Nye (2016, 46) writes, “The term ‘cyber deterrence’ can be confusing because theorists tend to focus on in-kind or in-domain deterrence rather than on a broad range of tools that can be used both actively and passively and with graduated effects.” Denning (2015, 11–15) sees the concept of cyber deterrence as problematic because it is so broad. According to Denning, one possible approach is to focus on classes of cyber-weapons and activities in cyberspace.

If we think that cyber-attacks are always aimed directly at information and indirectly at users of information, one can agree with Nye’s views: weapon systems in all types of weapons and their users need the right information. By blocking (from people or weapons) access to this information or making it unreliable, it is possible to influence from cyber domain to the other domains. As with other domains, it is possible to influence a cyber domain, for example, by destroying parts of psycho-physical layer.

Tor (2015, 107–108) has listed five basic principles of the cumulative deterrence paradigm that must be adapted to the cyber domain. They are as follows:

- A strategic message regarding the ‘red lines’ of the deterring party;
- An ability and willingness to carry out attacks on rivals, and to set a ‘price tag’ on cyberattacks;
- An ability and willingness to threaten and demonstrate capabilities both in cyber and through military tools or through diplomatic and financial tools;
- Overwhelming supremacy in cyberspace; and
- The effort to build a more robust and secure cyber infrastructure through constant development of technology and protection methods.

Achieving overwhelming supremacy in cyberspace is quite unrealistic due to the complexity of cyberspace (JP-12, 2018, I-2). Maintaining local, demarcated A2/AD zones can also be a challenge, although Russia, for example, is pursuing this with its RuNet program (Kukkola, 2020).

There are also problems with setting a red line and its credible defence, for example, in the clear communication of its existence and the consequences of its violation. The adversary may also rush to test the extent of the red line or try to stretch its boundaries by hybrid methods. Libicki (2012, 68) discusses about “salami tactics”, by which he means that once a state tolerates small violations, the cumulative effect begins to pinch, and the state realizes that it has neglected to establish a clear line to demarcate tolerable from intolerable violations and that can lead to an unpredictable set of results. Freedman (2020, 9) argues that: “Deterrence works best with unambiguous red lines, established over time, linked with vital interests, and backed by clear and credible messages, reinforced by known capabilities, about what will happen if they are crossed.” Hybrid influence, and cyber operations as part of it, is precisely a matter of testing the red lines of an adversary, creating uncertainty in political or official decision-making or applications of law, and thus the possibility of using military retaliation.

Challenges related to the build of cumulative cyber deterrence include, for example, its credibility, the communication of the cost of the attack, psychological influence on the adversary, and problems of manipulation and regulation. A credible cumulative cyber deterrent requires that the attack is also responded to or retaliated against. Bradly (2018, 45) argues that most retaliations are not feasible in cyberspace because they require the identification of the perpetrator, retaliation within a proximate temporal range, and proportionality while a state must also possess a specific cyber weapon system tailored to its target.

3. The Cumulative cyber deterrence in the military concepts of Russia and Israel

3.1 Russia

In the Russian philosophy of warfare, concealment and deception play an important role. General Gerasimov, the chief of staff of the Russian armed forces, has stressed (Krasnaya Zvezda, 4 March 2019) the importance of

considering modern warfare as consisting of military and non-military means of war, along with the achievement of surprise. He also highlighted the importance of preventive measures, the identification of vulnerabilities, creating deterrence, and maintaining the ability to take strategic initiative.

Russian Deputy Prime Minister Dmitry Rogozin (Rogozin, 2013) said already in 2013 that enemy can paralyze critical infrastructure of a target state with cyber-attacks). According to Russian experts Stuxnet was the first example of the cyber warfare and such an attack on Russian targets could cause enormous damage to Russia's economy (Orlov, 2011).

Putin has stated that the Soviet Union was a besieged fortress and under threat of attack from the West (Aron, 2008). As a continuation of this the Russian leadership has created a narrative according which Russia is under siege and under threat of attack from the West. The narrative that Russia is target of enemy states (Facon, 2016) and the country's perceived geostrategic and technological vulnerability (Covington, 2016), combined with Russia's feeling of a hostile world (Facon, 2017), have strengthened the Russian narrative of the besieged fortress (Igumnova, 2011). As a besieged fortress, Russia needs to be protected.

This besiege has expanded to cyberspace. According to Igor Ashmanov, a Russian ICT specialist, the cyber warfare against Russia is waged every day and no rules of war apply to it (Yarovaya, 2013). As a part of this warfare Russia has actively built its cyber deterrence. The main tasks of the Russian Federation regarding deterring and preventing military conflicts is to create conditions to reduce the risk of using information and communications technologies for military-political purposes (National Security Strategy 2021, art. 40).

Russia's way of building deterrence can be explained by the factors and elements of Russian strategic culture. The main factors influencing Russian strategic culture are its sense of vulnerability, fear of surprise attack, the narrative of Russia as a besieged fortress and the concept of permanent war. Russia is lagging behind the leading countries in the development of competitive information technology, and this gap strengthens the Russian perception of its strategic vulnerability in cyberspace. Russia's cyber deterrence consists of improved protection of its critical information infrastructure, preparations to isolate the Russian segment of the Internet from the global Internet, intensified surveillance, the ban of user anonymity on RuNet, and the aspiration to replace imported information and communication technology with Russian-produced ICT (Kari, 2019).

Russia tries to maintain the status quo or change it, not just threatening the adversary with retaliation or by denying it of its objectives through intolerable risks and costs (Kukkola, 2020, 185). Russia also shapes the strategic operating environment and manipulates its adversaries' perception while influencing its strategic behavior and using different flexible non-linear and cross-domain methods (Adamsky, 2021, 161, 170, 172).

Russia operates in the cyber environment. In addition to the cyber espionage targeted at the US Democratic Party in 2016, Russia was credited with intrusion into the e-mails of representatives of the Norwegian and Finnish parliaments in 2020, all of which Russia has actively denied. Russia has been linked to violent episodes such as that around the Bronze Soldier in Estonia in 2004, the Georgian War in 2008 and the conquest of Crimea in 2014. All these conflicts have been preceded by a cyber-attack on some socially important system in the target country. On the basis of these examples, it can be estimated that Russia and its authorities have good opportunities to control cyber-activities within its sphere of influence.

3.2 Israel

The public defense doctrine of the Israel Defense Forces (IDF, 2015), or the so-called Eizenkot's IDF Strategy, emphasizes strategic and tactical deterrence via cyberwarfare (Frei, 2020, 9). The strategy notes that deterrence against any enemy must be generalized and cumulative over time in order to maintain the existing situation and frame "rules of the game" favorable to Israel. Israel's Cyber Security Strategy (2017) also highlights the importance of operations and active efforts to confront the sources of the threats. The concept of operations defines three operational layers: Aggregate Cyber Robustness, Systemic Cyber Resilience and National Cyber Defense.

The development of capabilities and building up forces are based on strengthening strategic and tactical deterrence via cyber warfare (Belfer Center, 2016, 24, 48). Tor (2015, 111–112) argues that the cumulative deterrence theory developed in Israel in a conventional strategic context. Deterrence is perceived as a spectrum rather than a dichotomous, binary state. "The cumulative deterrence paradigm considers sporadic, short bursts

of violence as an integral part of a 'learning process' between the opposing parties. Such intermittent strategic interactions are meant to lead the deterred party to understand the 'red lines' of the deterring party."

Israel's concept of cumulative cyber deterrence is in nature deterrence by punishment. That includes limited and designed deterrence and influence operations, which are typically combined with a limited use of conventional force. The purpose of the operations is to reinforce the deterrent by trying to influence the adversary's behavior and to set the rules of the game for hostile interaction (Tor, 2015, 105–107). According to Almog (2004, 3, 6), classical deterrence and cumulative deterrence differs from conceptualization and implementation to desired results. On the macro level, cumulative deterrence seeks to create an image of overwhelming military supremacy and on the micro level, it relies on specific military responses to specific threats or hostile acts.

Maintaining deterrence and its credibility requires constant renewal. This means that the response to attacks should be immediate, certain, and the amount of force properly calibrated to the attack (Almog, 2004, 6; Shamir, 2020, 275, 276). Iran and Hamas have often been behind the significant public cyber-attacks against Israel. Israel has retaliated against the attacks in, for example, the 2010 Stuxnet attack against Iran's nuclear power plant and the retaliation in 2019 for Hamas's offensive cyber-attacks they bombed a Hamas cyber center (Frei, 2020, 6).

In assessing the concept of Israel's cumulative cyber deterrence, it is good to take into account geopolitical, political, historical, cultural, and diplomatic considerations. A similar concept that flexibly combines cyber and conventional methods of using force would not necessarily work in the Western countries. First, the use of military force has been a part of "everyday life" (on their own soil and borders) and a necessary activity throughout Israel's history due to its location and the presence of a hostile minority. Second, in no way is, from the point of view of the actors in the area referred to above, the use of military force in any way surprising or exceptional, but part of an ongoing conflict. Third, Israel's strong alliance with the United States helps strengthen the legitimacy of its actions from the perspective of international law. Thus, it can be estimated that Israel's cumulative cyber deterrence alone does not work, but contributes to strengthening Israel's overall deterrence. The problem with the Israeli concept of cumulative cyber deterrence is that while they seek to draw red lines and respond to the crossing of them with conventional force if necessary, the adversary nonetheless seeks to defy their defenses over and over again. It is no longer a question of the effects of cyber deterrence or its failure, but of the state of mind of the adversary.

4. Conclusions

The cumulative cyber deterrence in order to operate independently requires a limited and defined cyber environment, one that is either formed in the operating environment of the deterrent creator A2/AD or alternatively to access an adversary's A2/AD. Second, in order to function as their own deterrent system, building a cyber deterrent requires a large number of different cyber weapons and attack methods because they are disposable against the same adversary. However, due to the uniqueness of the cyber environment as a combat domain, the diversity of change factors, it may be more realistic to see the cumulative cyber deterrence as part of the overall deterrence of states.

The cyber deterrence can be seen as consisting of the strategical, operational, tactical and technological capabilities of the state or its allies. The cumulative cyber deterrence presupposes that its existence, means, and the will to implement it have been credibly communicated through either open or tacit strategic communication to the deterrent object, without the deterrent object receiving information that can circumvent the methods that maintain the deterrent. Second, the functioning of cyber deterrence as an independent concept requires that the adversary does not have significant conventional weapons available or has no will (e.g. political or legal or technological barriers) to use them.

The comparison between the cumulative cyber deterrence concepts of Russia and Israel showed that Russia has the time and resources to wait, shape the cyber domain and signal potential adversaries about consequences. Israel's concept of cumulative deterrence is more straightforward and emphasizes the need for an immediate and credible answer. Israel's strategic deterrence can be described as deterrence by retaliation, while Russia's strategic deterrence, especially its cyber deterrence, is more deterrence by denial in nature, although the means they use are often offensive.

The concept of cumulative cyber deterrence depends on the other means of deterrence (political and military will, plans, conventional military capabilities, technological skills) available to the state. Both Israel and Russia have all these qualities. There are several similarities in the military thinking based on strategic culture behind the defense doctrines of Russia and Israel. In particular, the perception of an ongoing and permanent war in the cyber domain distinguishes these states and their concept of cumulative deterrence from "Western" states. The overall deterrence strategy of these states is based on their strategic culture and their perception of threats, which differ from each other. A key difference in the deterrence strategies of these states is that Israel uses cumulative methods to make it clear where the red lines go, while Russia's strategic goal is to blur them and create the fog of war.

In the scope of the future, it is obvious that both Israel and Russia are investing in the development of artificial intelligence, robotics and autonomous weapon systems. Success in this development is likely to reinforce deterrence, but on the other hand, this also highlights the importance of cyber defense: advanced weapons systems can be also unpredictable and vulnerable.

References

- Adamsky, D. (2021) Deterrence à la Ruse: Its Uniqueness, Sources and Implications. F. Osinga and T. Sweijts (eds.), NL ARMS Netherlands Annual Review of Military Studies 2020, NL ARMS, https://doi.org/10.1007/978-94-6265-419-8_9. pp.161-175.
- Almog, D. (2004) Cumulative Deterrence and the War on Terrorism .United States Army War College Quarterly (Winter 2004-05), v.34 no.4, p.4-19. <https://www.hsdl.org/?view&did=453973>.
- Aron, L. (2008) 'The Problematic Pages. In memory of Alexander Solzhenitsyn', The New Republic. (24 September 2008). <https://newrepublic.com/article/62070/the-problematic-pages>.
- Arquilla, J. & Ronfeld, D. (1993) Cyberwar is coming! <https://www.rand.org/pubs/reprints/RP223.html>.
- Belfer Center for Science and International Affairs (2016) Deterring Terror. How Israel Confronts the Next Generation of Threats. English Translation of the Official Strategy of the Israel Defense Forces Foreword by Graham Allison. <https://www.belfercenter.org/sites/default/files/legacy/files/IDFDoctrineTranslation.pdf>.
- Brantly, A. (2018). The cyber deterrence problem. In T. Minárik, R. Jakschis, & L. Lindström (Eds.), *10th international conference on cyber conflict*. Tallinn: NATO CCD COE Publications. Retrieved from: <https://ccdcoe.org/uploads/2018/10/Art-02-The-Cyber-Deterrence-Problem.pdf>.
- Covington, S. (2016) 'The Culture of Strategic Thought Behind Russia's Modern Approaches to Warfare.' Belfer Center. Harvard Kennedy School. <https://www.belfercenter.org/sites/default/files/files/publication/Culture%20of%20Strategic%20Thought%203.pdf>.
- Denning, D.E. (2015) Rethinking the Cyber Domain and Deterrence. https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-77/jfq-77_8-15_Denning.pdf.
- Department of Defense Science Board (2017) Task Force on Cyber Deterrence. <https://www.hsdl.org/?collection&id=2724>.
- Facon, I. (2016) 'Russian Strategic Culture in the 21st Century: Redefining the West-East Balance', in Tellis A, Szalwinski A and Wills M (eds) *Understanding Strategic Cultures in the Asia-Pacific, Strategic Asia 2016-2017*, The National Bureau of Asian Research, 62-89. http://nbr.org/publications/strategic_asia/pdf/SA16_ExecutiveBrief.pdf.
- Facon, I. (2017) 'Russia's national security strategy and military doctrine and their implications for the EU', European Parliament's Sub-Committee on Security and Defence. [http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/578016/EXPO_IDA\(2017\)578016_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/578016/EXPO_IDA(2017)578016_EN.pdf).
- Freedman, L. (2020) Introduction—The Evolution of Deterrence Strategy and Research. In Osinga, F. and Sweijts, T. (Eds.): *Deterrence in the 21st Century—Insights from Theory and Practice*. NL ARMS Netherlands. Annual Review of Military Studies 2020. <https://link.springer.com/content/pdf/10.1007%2F978-94-6265-419-8.pdf>. pp.1-10.
- Frei, J. (2020) Israel's National Cybersecurity and Cyberdefence Posture. Policy and Organizations. css.ethz.ch/en/publications/risk-and-resilience-reports.html.
- Jasper, S. (2018) U.S. Strategic Cyber Deterrence Options. http://centaur.reading.ac.uk/79976/1/22839264_Jasper_thesis.pdf.
- State of Israel (2017) Israel National Cyber Security Strategy in Brief. https://cyber.haifa.ac.il/images/pdf/cyber_english_A5_final.pdf.
- Johnston, A. (1995). *Cultural Realism: Strategic Culture and Grand Strategy in Chinese History*. Princeton University Press 1995.
- Joint Chief of Staff (2018) JP 3-0, Joint Operations. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_0ch1.pdf?ver=2018-11-27-160457-910.
- Joint Chief of Staff (2018) JP-3-12, Cyberspace Operations. https://irp.fas.org/doddir/dod/jp3_12.pdf.
- Igumnova, L. (2011) 'Russia's Strategic Culture Between American and European Worldviews', *The Journal of Slavic Military Studies*, Volume 24. doi: 10.1080/13518046.2011.572729.
- Kari, M. J (2019) Russian Strategic Culture in Cyberspace Theory of Strategic Culture – a tool to Explain Russia's Cyber Threat Perception and Response to Cyber Threats. JYU DISSERTATIONS 122.

- Kasapoglu, C. (2015) Russia's Renewed Military Thinking: Non-Linear Warfare and Reflexive Control. NATO Defence College, Rome – No. 121.
<https://cco.ndu.edu/Portals/96/Documents/Articles/russia's%20renewed%20Military%20Thinking.pdf>.
- Krasnaya Zvezda (4.3.2019) Векторы развития военной стратегии. <http://redstar.ru/vektory-razvitiya-voennoj-strategii/>.
- Kukkola, J. (2020) Digital Soviet Union. The Russian national segment of the Internet as a closed national network shaped by strategic cultural ideas. National Defence University. Series 1: Research Publications No. 40.
https://www.doria.fi/bitstream/handle/10024/177157/Kukkola_Digital%20Soviet%20Union_finalnet.pdf?sequence=3&isAllowed=y.
- Libicki, M. C. (2012) Crisis and escalation in cyberspace. (Santa Monica: RAND Corporation)
<https://www.rand.org/pubs/monographs/MG1215.html>.
- Nye, J. S. Jr. (2016): Deterrence and Dissuasion in Cyberspace. Quarterly Journal: International Security 2016/17) pp.44-71
https://www.belfercenter.org/sites/default/files/files/publication/isec_a_00266.pdf.
- Orlov, V. (2011) Start of new battles, Moskovskie Novosti. (21 April 2011). <http://www.mn.ru/newspaper/world/68636>.
- Raitasalo, J. – Sipilä, J. (2008) "Näkökulmia sotaan" Sota – teoria ja todellisuus. Näkökulmia sodan muutokseen. pp. 1-10.
<http://urn.fi/URN:ISBN:978-951-25-1894-4>.
- Rogozin, D. (2013) Speech by Dmitry Rogozin at a press conference in the "RG" (28 June 2013). (in Russian)
<https://rg.ru/2013/06/28/doklad.html>.
- Shamir, E. (2020) Deterring Violent Non-state Actors, 263-286. In Osinga, F. and Sweijs, T. (Eds.): Deterrence in the 21st Century—Insights from Theory and Practice. NL ARMS Netherlands. Annual Review of Military Studies 2020.
<https://link.springer.com/content/pdf/10.1007%2F978-94-6265-419-8.pdf>. pp. 263-286.
- Schulze, M. (2020) Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations
https://ccdcoe.org/uploads/2020/05/CyCon_2020_10_Schulze.pdf.
- Soesanto, S. and Smeets, M. (2020) Cyber Deterrence: The Past, Present, and Future. In Deterrence in the 21st Century—Insights from Theory and Practice. Annual Review of Military Studies 2020. pp. 386-397
<https://link.springer.com/content/pdf/10.1007%2F978-94-6265-419-8.pdf>.
- Thomas, T. (2004) Russia's Reflexive Control. Theory and the Military. Journal of Slavic Military Studies 17: 237–256.
https://www.rit.edu/~w-cmmc/literature/Thomas_2004.pdf.
- Tor, U. (2015) Cumulative Deterrence' as a New Paradigm for Cyber Deterrence. December 2015 Journal of Strategic Studies 40(1-2):1-26
- Williams, B. D. (2017) Meet the scholar challenging the cyber deterrence paradigm.
<https://www.fifthdomain.com/home/2017/07/19/meet-the-scholar-challenging-the-cyber-deterrence-paradigm/>.
- Yarovaya, M. (2013) 'Igor Ashmanov: 'Today information domination is the same as air superiority'',
<https://ain.ua/2013/05/01/igor-ashmanov-segodnya-informacionnoe-dominirovanie-eto-vse-ravno-cto-gospodstvo-v-vozduxe>.

Masters Research Papers

Performance Implications for Multi-Core RISC-V Systems with Dedicated Security Hardware

Samuel Chadwick, Scott Graham and James Dean

Air Force Institute of Technology, Wright-Patterson Air Force Base, USA

samuel.chadwick@afit.edu

scott.graham@afit.edu

james.dean@afit.edu

Abstract: The RISC-V instruction set architecture (ISA) is a promising open-source architecture supporting the Open Era of Computing. As RISC-V matures, consumers, industry leaders, and nation states are looking at the potential benefits RISC-V offers —especially for secure systems which may require privileged architecture implementations, physical memory protection (PMP), or trusted execution environments (TEEs) among other hardware-based security primitives. The inclusion of these security technologies unavoidably impacts the performance of any given compute system. To quantify the performance impacts introduced by secure enclave processing, representative computational benchmarks are executed on the Freedom U74-MC System-on-a-Chip (SoC) onboard the HiFive Unmatched development board by SiFive. These benchmarks are conducted across applicable modes of the RISC-V Privileged ISA specification to analyze Privileged ISA and PMP performance implications for Confidential Computing. To evaluate performance impacts, a theoretical model is applied to represent the interactions of the security monitor. The Keystone enclave framework tasks the security monitor with enforcing strict adherence to system security primitives while the Phoronix Test Suite (PTS) captures performance data. Individual benchmarks are conducted both with and without secure enclave technologies to characterize representative performance metrics.

Keywords: Confidential Computing, Keystone Security Monitor, Performance Characterization, Physical Memory Protection, RISC-V, Secure Enclave, Trusted Execution Environment

1. Introduction

1.1 Motivation

Existing encryption techniques provide confidentiality, integrity, and availability for *data at rest* and for *data in transit* yet are rarely employed to protect *data in use* during execution. Confidential Computing performs computation within hardware-based Trusted Execution Environments (TEEs) to adequately protect *data in use*. TEEs ensure that compute systems maintain confidentiality, integrity, and availability by implementing security primitives to provide secure boot, a secure source of randomness, and enable remote attestation. While proprietary TEE implementations exist —such as Intel Software Guard eXtensions (SGX), Advanced RISC Machines (ARM) TrustZone, and Advanced Micro Devices (AMD) Secure Encrypted Virtualization— they each impose threat model limitations and restrict changes to their underlying architectural Intellectual Property (IP). Keystone Enclave is an extensible open-source TEE alternative which leverages the Privileged RISC-V Instruction Set Architecture (ISA), Physical Memory Protection (PMP), and other RISC-V Security Primitives to enforce Confidential Computing. Keystone Enclave reduces the Trusted Computing Base (TCB) by applying a software-defined and hardware-enforced framework for developing custom threat models.

1.2 Hypothesis

Confidential Computing unavoidably impacts system performance.

1.3 Approach

To appropriately evaluate architectural design implementations, a functional closed queueing network is applied to measure average system response times for secure enclave applications running on RISC-V systems. Performance impacts are characterized by applying statistical methodologies to ensure that compute systems fulfil operational requirements while simultaneously enforcing security obligations.

Leveraging the HiFive Unmatched as the system under test, we evaluate representative workstation workloads via benchmarking: First with an unmodified Linux system without the Keystone Enclave; Then with the Keystone Enclave installed, but with the SM features unused; And finally, with the Keystone Enclave installed, with varying instances of secure enclaves. Each test records the execution time and assigns benchmark scores accordingly. To ensure accurate statistical and analytical analyses, we perform automated benchmark scheduling to dynamically repeat tests until our results fall within a predefined standard deviation. We then evaluate collected

data under the applied modelling framework to characterize trade-offs between overall system performance and security.

The next sections within this paper describe the following information: Section II provides relevant background knowledge; Section III describes experimental design configurations; Section IV describes preliminary benchmark results; while Section V concludes with current and future contributions.

2. Background

2.1 Modelling framework efficacy

Gorbachov et al. (2019) modelled embedded system performance impacts by applying a subject-object system framework to the simple closed queue network model. The active component responsible for managing and controlling access of subjects to objects is broadly described as a Reference Monitor (Gorbachov et al., 2019). As implemented, Keystone applies this theoretical model by leveraging RISC-V security primitives (such as the privileged ISA modes and PMP capabilities) to construct customizable TEEs (Lee, et al., 2020) and rebrands the previously described Reference Monitor subject as the Security Monitor (SM). Recent research has demonstrated the non-triviality for creation, execution, and destruction of secure enclaves for RISC-V systems and applications (Tullos et al., 2020). This work extends prior study by maturing the system under test from Field Programmable Gate Array (FPGA) technology to dedicated Application Specific Integrated Circuits (ASICs) found in commodity personal computer (PC) hardware.

The modified functional closed queuing network model, transposed for Keystone SM by Tullos, et al. (2021) is shown in Figure 1. This transposition characterizes secure enclave *creation* latency to correlate performance metrics with secure execution. From this model, the system response time R can be measured and regarded as performance overhead for tasks which require enclave initialization.

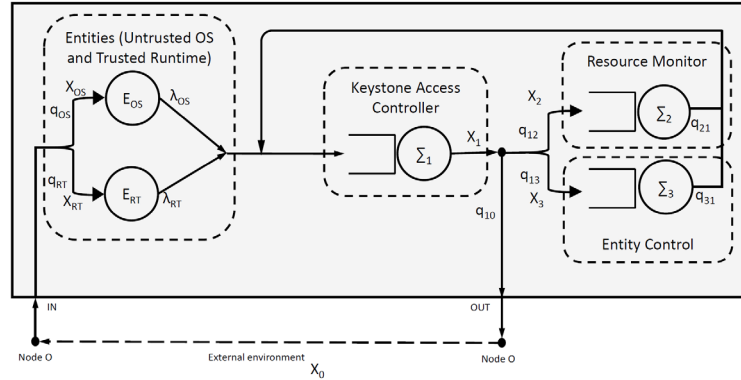


Figure 1: Functional closed queuing network model for Keystone SM enclave creation (Tullos, et al., 2021)

As proposed by Gorbachov et al. (2019) and as applied by Tullos et al. (2021), system service time R is based on the system output rate X_0 where the system flow-balance equations are as follows:

$$\begin{aligned} X_0 &= q_{10} \cdot X_1 \\ X_1 &= X_0 + X_2 + X_3 \\ X_2 &= q_{12} \cdot X_1 \\ X_3 &= q_{13} \cdot X_1 \end{aligned}$$

The Interactive Response Time Formula presented by Gorbachov et al. (2019) is applied to calculate the mean system service time R , where N represents the number of subjects (secure processes), and λ describes the rate of requests for objects (secure accesses):

$$R = \frac{N}{X_0} - \frac{1}{\lambda}$$

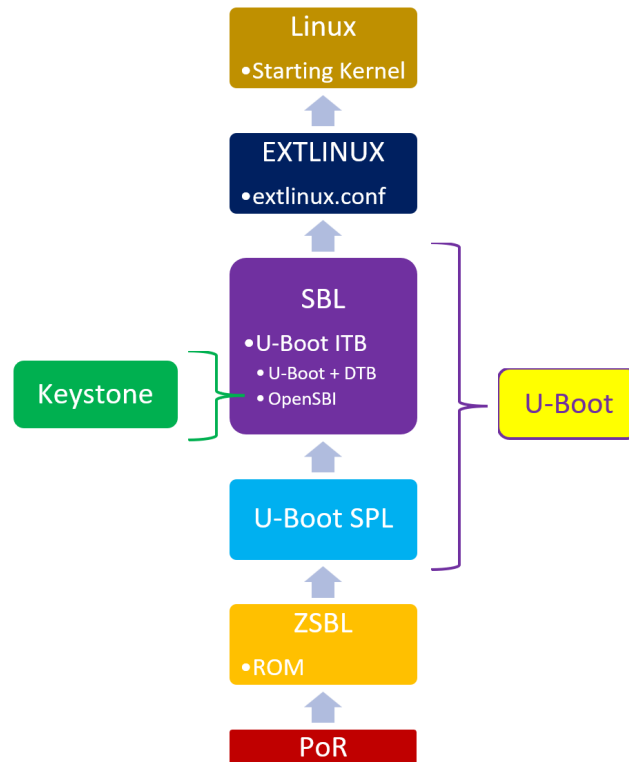
It follows that system service time R for enclaved computing can be determined by measuring the throughput of either the subject reference monitor Σ_2 , or the object reference monitor Σ_3 ; these throughputs are labelled X_1 , X_2 , or X_3 accordingly. With known routing frequencies for q_{12} or q_{13} , the modified system flow-balance equations are solved and the system output rate X_0 is applied to calculate system service time R ; assuming the remaining N and λ variables are known. For our computing system, the number of secure processes and number

of secure accesses are deterministic and based on the number of enclave processes and the processor frequency.

2.2 Keystone: an open-source secure enclave framework for RISC-V processors

The Keystone Enclave is designed for and built upon the RISC-V ISA. As an open-source, customizable TEE framework, it supports the four privilege modes outlined by “The RISC-V Instruction Set Manual, Volume II: Privileged Architecture:” U-mode (user) for applications, S-Mode (supervisor) for the kernel, H-mode (hypervisor) for hypervisor-level isolation, and M-mode (machine) for direct access to physical system resources. Keystone leverages RISC-V isolation primitives such as privilege modes and the PMP standard to enable isolation of memory-mapped control features *below* untrusted code (Lee et al., 2020).

We use the standardized boot flow, detailed in the “SiFive FU740-C000 Manual,” for our system under test. This boot flow executes in the following order: (1) Power on Reset (PoR); (2) Zeroth Stage Bootloader (ZSBL) stored within Read Only Memory (ROM); (3) U-Boot Secondary Program Loader (SPL); (4) Second Bootloader (SBL) supporting (i) the U-Boot Image Tree Blob (ITB), (ii) Device Tree Blob (DTB) and (iii) OpenSBI; (5) EXTLINUX containing extlinux.conf; and (6) the Linux Kernel. Keystone operates on standard RISC-V cores and requires a hardware root of trust. Installation on the system is accomplished by installing Keystone as M-mode bootloader software (i.e., firmware) using an *out-of-tree* platform build configuration supported by OpenSBI. Figure 2 captures the described boot flow and provides context for how Keystone enforces resource accesses between secure and unsecure processes.



Once installed, the Keystone SM facilitates OS secure enclave requests by calling the appropriate Keystone Supervisor Binary Interface (SBI) functions described in Table 1. Keystone enclaves experience three distinct phases during their lifecycle: creation, execution, and destruction. Upon a *creation* request by the OS E_{OS} , the Keystone access controller Σ_1 measures the enclave memory, ensuring that the OS E_{OS} correctly loaded the enclave binaries into physical memory. Then, the access controller Σ_1 hashes the page contents and the virtual addresses along with configuration data to send to the Entity Monitor (EM) Σ_3 . At *execution*, the access controller Σ_1 sets PMP entries and delegates control to the Resource Monitor (RM) Σ_2 . After program completion, the runtime entity E_{RT} calls the *exit* function prompting the OS E_{OS} to initiate a *destruction* request. Upon a *destruction* call, the access controller Σ_1 clears the enclave memory region prior to returning the memory address space to the OS E_{OS} (Lee et al., 2020).

Table 1: Keystone SM SBI Functions

Caller	SBI Function	Description
E_{OS}	create	Validate & measure the enclave
	run	Start enclave & boot runtime
	resume	Resume enclave execution
	destroy	Clean & release enclave memory
E_{RT}	stop	Pause enclave execution
	exit	Terminate the enclave
	attest	Get signed attestation report
	random	Get secure random values

Due to the rapid adoption of the RISC-V ISA, expanding accessibility to RISC-V hardware, and maturing application development landscape, Keystone itself has undergone its first major version release. Keystone Enclave 1.0.0 (released on 2 March 2021) exclusively supports the RISC-V Open Source Serial Binary Interface (OpenSBI) which replaces the deprecated Berkley Boot Loader (BBL) to enforce a standardized RISC-V Boot Flow. Such changes offer opportunities to prove initial performance conclusions atop wider ranges of representative test platforms and use cases.

2.3 HiFive Unmatched by SiFive

We selected the HiFive Unmatched RISC-V Powered Linux Development Kit as our test platform of choice for four primary reasons: (a) form factor, (b) commodity hardware compatibility, (c) Linux support, and (d) enhanced System-on-a-Chip (SoC) monitoring. At the time of writing, the HiFive Unmatched is the only available RISC-V platform which supports the mini-Information Technology eXtended (ITX) form factor used by many x86_64 ISA systems. SiFive’s decision to support this form factor enables flexible use of Peripheral Component Interconnect Express (PCIe) and Non-Volatile Memory Express (NVMe) hardware commonly found within the PC ecosystem. In practice, the native ability to run Ubuntu 21.04 (Hirsute Hippo) as the host OS simplifies testing procedures and provides a realistic environment suitable for benchmarking system performance. Aside from the form factor, hardware compatibility, and familiar OS, the Unmatched lends a unique perspective to TEE performance characterizations because it dedicates a fifth Freedom S7 monitor core exclusively for monitoring the four Freedom U74 cores. Precise SoC specifications are found in the FU740-C000 Manual, with our test configuration explicitly detailed in Section 3.

While other ASIC RISC-V systems offer similar or competing features, at present no other commercially available offering culminates all these features onto a single package. Moreover, company leadership at SiFive is comprised of three RISC-V ISA co-founders: Yunsup Lee, Chief Technology Officer; Krste Asanovic, Chief Architect, and Andrew Waterman, Chief Engineer. Their involvement at SiFive inspires confidence for long-term hardware support of SiFive products as RISC-V specifications mature.

2.4 Related work

Executing software with guaranteed integrity and confidentiality on a remote computing system, which is owned and maintained by an untrusted party, requires secure remote computation. Intel’s Software Guard Extensions (SGX) aims to provide a trusted computing solution for x86 platforms by leveraging trusted hardware within the remote computer (Costan et al., 2017). Performance loss estimation studies have been explored which attempt to correlate SGX protections with performance degradation; however, the complex overlapping between security and performance factors is not bijective (Zheng et al., 2018). Accordingly, performance characterizations are often limited to time intensive trial-and-error analyses (Weichbrodt, et al., 2018).

Tullos et al. (2021) extended the Reference Monitor framework proposed by Gorbachov et al. (2019) to measure Keystone’s system performance impact on FPGA softcore processor implementations. The applied Keystone SM framework describes the object reference monitor as the Resource Monitor (RM) and the subject reference monitor as the Entity Monitor (EM). The RM is responsible for ensuring system resource accesses adhere to privilege modes and PMP configurations, while the EM monitors and enforces authorized enclave usage by ensuring that TEEs are created, executed, and destroyed as authorized (Tullos et al., 2021). Figure 3 illustrates the adapted SM concept model and captures its conceptual operation. Our work supplements this Keystone SM framework (depicted in Figure 1) by comparably evaluating system performance impacts on RISC-V ASIC hardware across the enclave lifecycle –throughout *creation*, *execution*, and *destruction*.

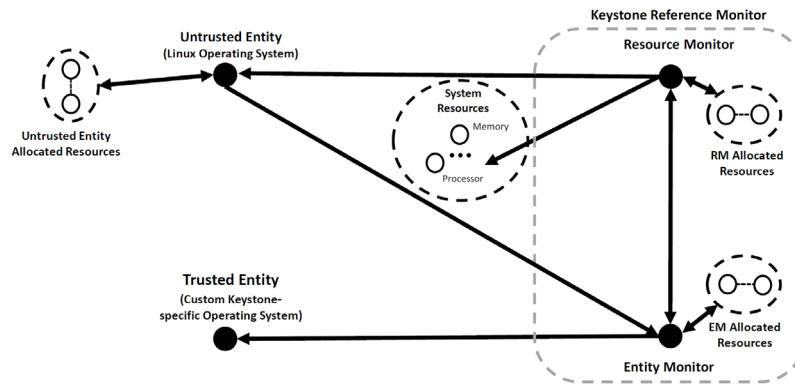


Figure 3: Keystone SM Concept Model (Tullos, et al., 2021)

3. Experimental design

3.1 Test platform configuration

The HiFive Unmatched SoC (FU740-C000) includes a 64-bit S7 monitor core and four 64-bit U74 application cores. The S7 monitor core has a dual-issue in-order execution pipeline supporting a peak sustained execution rate of two instructions per clock cycle (IPC). The monitor core supports M and U privilege modes for standard Multiply, Atomic, and Compressed RISC-V extensions (RV64IMAC). Each of the U74 application cores have dual-issue in-order execution pipelines supporting a peak sustained execution rate of two IPC per core. The application cores support M, S, and U privilege modes for standard Multiply, Single-Precision Floating Point, Double-Precision Floating Point, Atomic, and Compressed RISC-V extensions (RV64IMAFDC or RV64GC). Only the U74 application cores are Linux capable, although all five cores maintain a fully-coherent two Megabyte (2MB) shared level two (L2) cache. Assembly level software development for the S7 monitor core exceeds the scope of this proposal, although future exploration of its interoperability and monitoring capabilities warrants future study.

Our test system makes use of the NMVe M.2 2280 and 2230 form factors to utilize a Samsung 980 PRO PCIe 4.0 NVMe M.2 Solid-State Drive (SSD) for booting and an Intel Wireless-AC 9260 Wi-Fi card for internet access. Figure 4 captures our system under test as configured. The x16 PCIe Gen 3 Expansion Slot remains unused but can be configured to support select Advanced Micro Devices (AMD) RX 500-series or Radeon HD 6000-series graphical processing units (GPUs). Graphical capabilities and implications are left for future investigation.

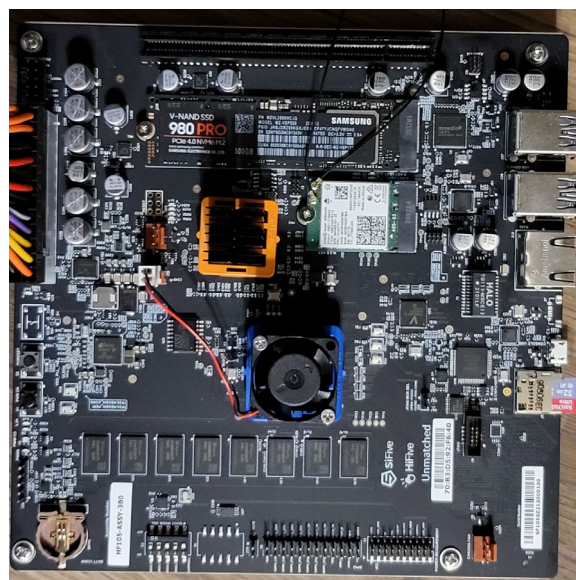


Figure 4: HiFive Unmatched Development Platform (test configuration with SSD and Wi-Fi cards installed)

Officially supported Linux distributions for RISC-V and for the Unmatched are currently limited. The distribution included on the provided microSD card equips users with an OpenEmbedded implementation; however, in order to better capture common usage scenarios, we selected Ubuntu 21.04 (Hirsute Hippo) as our host OS due to its

popularity and ample community support. Thanks to official support by Canonical, many conveniences of the existing Ubuntu ecosystem and the Debian Linux distribution are readily available in support of our specific hardware configuration –support for our desired software evaluation programs notwithstanding.

3.2 Benchmark suite configuration

We selected the Phoronix Test Suite (PTS) as our benchmarking platform during testing. The PTS offers a comprehensive testing and benchmarking experience while providing an extensible framework for adding custom benchmarking tests. From the Ubuntu terminal, the PTS enables effective, reproducible, and automated testing with an integrated remote management system to schedule, install, and archive system, test, and installation logs. Benchmark results are further configurable to automatically record to the OpenBenchmarking.org web viewer.

If desired, the PTS offers extensible markup language (XML) architecture the ability to run custom benchmarks and suites. Once added and installed, the PTS can be configured to automatically re-run tests if the results exceed a predefined standard deviation threshold. Without the pertinent need for a custom testing suite, we opted to generate preliminary benchmark results by using existing suites hosted by OpenBenchmarking.org.

3.3 Baseline performance characterization

For experimentation, we selected a compatible subset of the “Stress-NG” benchmarking suite for its intentional exercise of physical computer subsystems and of OS-to-Kernel interfaces. Openly published by Canonical, this suite encompasses over 260 stress tests exercising Central Processing Units (CPUs) and virtual memory operations among other subsystems. The “Stress-NG” Linux stress tool has been configured to execute 20 commonly used benchmarks on our Unmatched system running Ubuntu 21.04 with the “5.11.0-1020-generic (riscv64)” Kernel. These benchmarks were selected primarily because of their compatibility with the RISC-V ISA and for their common usage across numerous applications. Table 2 lists each benchmark selected from the suite to be executed, along with a brief description.

Table 2: Preliminary Benchmarking Metrics for Baseline Performance

	Benchmark	Description
0	Stress-NG 1.4.0	Linux Stress Test
1	MMAP	Memory Map
2	NUMA	Non-Uniform Memory Access
3	MEMFD	Anonymous Kernel Memory Management
4	Atomic	Atomic Operations
5	Crypto	MD5, SHA-256, SHA-512, scrypt, NT, yescrypt
6	Malloc	Memory Allocation
7	Forking	CPU Forking
8	IO_uring	Asynchronous Input / Output
9	SENDFILE	Read/Write
10	CPU Cache	Cache Thrashing
11	CPU Stress	Integer, Multiply, Floating Point, & Double Precision Operations
12	Semaphores	Shared Resources
13	Matrix Math	Two- & Three-Dimensional Matrix Operations
14	Vector Math	128-Bit Vector Operations
15	Memory Copying	memcpy() Method Operation
16	Socket Activity	IPv4, TCP Congestion Control
17	Context Switching	Memory Clobbering
18	Glibc C String Functions	Glibc C String Functions
19	Glibc Qsort Functions	Glibc Qsort Functions
20	System V Message Passing	System V Message Passing

4. Preliminary results and expectations

4.1 Baseline performance characterization

The RISC-V ISA, while maturing, is still new; compatible hardware optimized applications are not yet common. Unsurprisingly, our system benchmark results confirm this reality. For context, current generation Raspberry Pi 400 series devices frequently score double that of our HiFive Unmatched system. As an embedded development device, the ability to run Linux is impressive; however, much work remains for RISC-V adoption within the

datacenter and among consumers. Figure 5 details our precise testing configuration while Figures 6, 7, and 8 highlight baseline performance metrics and result confidence.

OpenBenchmarking.org	Phoronix Test Suite 10.6.0
SiFive RISC-V (4 Cores)	Processor
SiFive HiFive Unmatched A00	Motherboard
16GB	Memory
Samsung SSD 980 PRO 500GB + 32GB SD32G	Disk
Intel-AC 9260	Network
Ubuntu 21.04	OS
5.11.0-1020-generic (riscv64)	Kernel
X Server 1.20.11	Display Server
GCC 10.3.0	Compiler
ext4	File-System

Figure 5: Testing Configuration for Performance Characterizations

The baseline benchmark results charted below highlight the novelty of the RISC-V ISA and of supporting hardware. Benchmark scores are listed in operations per second, with higher scores indicating better performance. For context, at a comparable price point of \$600, an ARM or x86 system would handily achieve scores in the millions compared to our hundreds or thousands.

211008-HFU-ICCWS-Baseline	
	HiFive Unmatched - Ubuntu 21.04 - Stress-NG - No Keystone
stress-ng: MMAP	1.40
stress-ng: NUMA	14.46
stress-ng: MEMFD	6.78
stress-ng: Atomic	46477.32
stress-ng: Crypto	75.13
stress-ng: Malloc	1419228.21
stress-ng: Forking	2203.43
stress-ng: IO_uring	2359.91
stress-ng: SENDFILE	5816.28
stress-ng: CPU Cache	13.21
stress-ng: CPU Stress	169.61
stress-ng: Semaphores	110475.28
stress-ng: Matrix Math	518.02
stress-ng: Vector Math	364.58
stress-ng: Memory Copying	28.68
stress-ng: Socket Activity	174.78
stress-ng: Context Switching	98326.92
stress-ng: Glibc C String Functions	14647.47
stress-ng: Glibc Qsort Data Sorting	4.83
stress-ng: System V Message Passing	242686.42

OpenBenchmarking.org

Figure 6: Baseline Performance Results, Raw

By charting our results on a graph scaled for mature ISAs and hardware, it becomes apparent that RISC-V is still new. Currently, only memory allocation operations are comparable to traditional workstation performance.

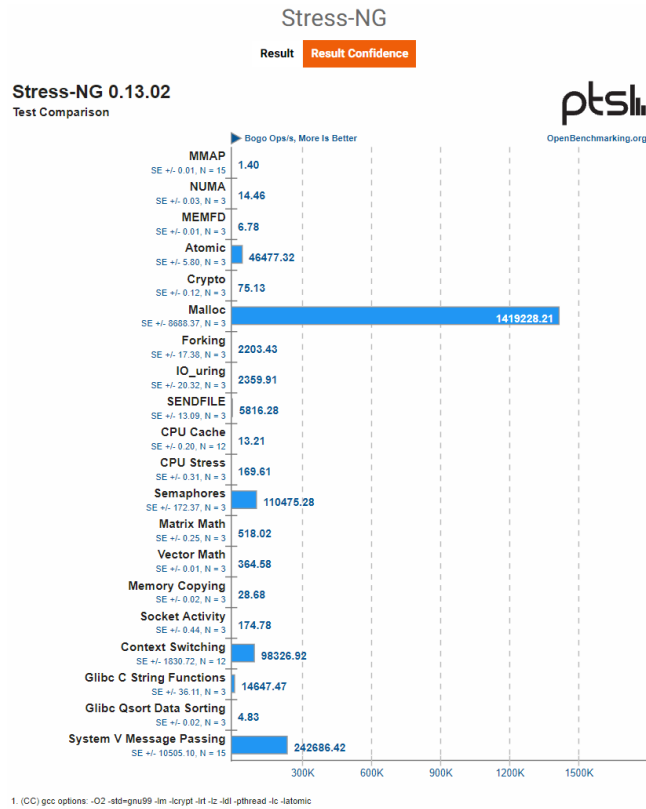


Figure 7: Baseline Performance Results, Charted

Confidence intervals for each benchmark are charted in Figure 8. and indicate that our results are consistent across repeated tests.



Figure 8: Baseline Performance Result Confidence, Charted

For further investigation, our complete benchmarking results are available online at the following Uniform Resource Locator (URL): <https://openbenchmarking.org/result/2110085-EALD-211008H69>.

5. Current and future work

5.1 Keystone performance characterization

The RISC-V security primitives applied by Keystone are unrivalled by any other framework. Unfortunately, at the time of writing, the required Keystone firmware configuration files are currently under development by the Keystone Enclave community to support the HiFive Unmatched. Enclave experimentation cannot be evaluated until either (a) a version of Keystone is released which supports our test hardware, or (b) we successfully implement these configuration files manually. The clear path forward is targeted firmware development, focused on enabling Keystone on the HiFive Unmatched while also promoting hardware-agnostic adoption.

5.2 Impacts of Keystone Enclave on system performance

Without a functional hardware implementation of the Keystone Enclave, concrete performance degradation conclusions have not yet been measured –especially for capturing performance implications related to context switching. Expected results will likely conclude that the performance overhead for enforcing Confidential Computing moderately degrades system performance. The cost of adding TEEs is significant; however, the degree to which degradation occurs during representative testing remains to be seen.

Disclaimer

The views expressed in this paper are those of the authors, and do not reflect the official policy or position of the United States Air Force, United States Space Force, Department of Defense, or the U.S. Government. This document has been approved for public release; distribution unlimited, case #88ABW-2021-0928.

References

- Asanović, K. and Patterson, D.A. (2014) *Instruction Sets Should Be Free: The Case for RISC-V*, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2014-146.
- Costan, V., Lebedev, I. and Devadas, S. (2016) *Sanctum: Minimal hardware extensions for strong software isolation*, In 25th USENIX Security Symposium (USENIX Security 16), pp. 857-874.
- Gorbachov, V., Batiaa, A.K., Ponomarenko, O. and Kotkova, O. (2019) *Impact Evaluation of Embedded Security Mechanisms on System Performance*, 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology, pp. 407-410.
- Lee, D., Kohlbrenner, D., Shinde, S., Song, D. and Asanovic, K. (2020) *Keystone: An Open Framework for Architecting Trusted Execution Environments*, In Proceedings of the Fifteenth European Conference on Computer Systems, pp. 1-16.
- SiFive FU740-C000 Manual v1p3*, p. 68. [online]. Available at: https://sifive.cdn.prismic.io/sifive/de1491e5-077c-461d-9605-e8a0ce57337d_fu740-c000-manual-v1p3.pdf (Accessed 10 October 2021)
- Tullos, J. C. (2021) *Characterizing Security Monitor and Embedded System Performance Across Distinct RISC-V IP-Cores*, AFIT-ENG-MS-21-M-087, p. 183. doi: 10.1109/TE.1962.4322266.
- Waterman, A., Lee, Y., Avizienis, R., Patterson, D.A. and Asanovic, K. (2021) *The RISC-V Instruction Set Manual Volume II: Privileged Architecture, Document Version 20210921*, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016- 129.
- Weichbrodt, N., Aublin, P.L. and Kapitza, R. (2018) *Sgx-perf: A performance analysis tool for Intel SGX enclaves*, Proceedings of the 19th International Middleware Conference, pp. 201-213.
- Zheng, W., Cao, S., Gao, Z., Wu, X. and Ding, Q., (2018) *The Performance Evaluation Model of Intel SGX-Based Data Protection*, 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, pp. 1289-1292.

Analysis of Image Thresholding Algorithms for Automated Machine Learning Training Data Generation

Tristan Creek and Barry E. Mullins

Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

Tristan.Creek@afit.edu

Barry.Mullins@afit.edu

Author note

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

Abstract: Secured compounds often safeguard physical layout details of both internal and external facilities, but these details are at risk due to the growing inclusion of Light Detection and Ranging (LiDAR) sensors in consumer off-the-shelf (COTS) technology such as cell phones. The ability to record detailed distance data with cell phones facilitates the production of high-quality three-dimensional scans in a discrete manner which directly threatens the security of private compounds. Therefore, it behooves the organizations in charge of private compounds to detect LiDAR activity. Many security cameras already detect LiDAR sources as generic light sources in specific conditions, but further analysis must identify these light sources as LiDAR sources in order to alert an organization of a potential security incident. Testing confirms the feasibility of identifying some LiDAR sources based on the color and intensity of light shined directly into a camera sensor, but this analysis proves inadequate for cell phone LiDAR. However, the unique intensity and pattern characteristics of cell phone LiDAR reflected off a surface can potentially be identified by an object identification machine learning model. In order to train a model to identify a LiDAR object, we must first produce a training dataset containing marked and labelled LiDAR objects. To do this, we apply an image thresholding algorithm to isolate the LiDAR object in an image to calculate its bounding box. The image thresholding algorithm directly affects the bounding box accuracy, so we test two different algorithms and find that Otsu's image thresholding algorithm performs best, resulting in 99.5% accurate bounding boxes.

Keywords: image thresholding, machine learning, LiDAR, CMOS

1. Introduction

The use of Light Detection and Ranging (LiDAR) devices spans across decades of history in various implementations for the military, private sector, and even civilian hobbyists (McManamon, Kamerman and Huffaker, 2010; Molebny, Kamerman and Steinvall, 2010). Although the basic principle of LiDAR relies on simply emitting infrared light at a surface or object and capturing what reflects back, the long-term development of LiDAR has produced countless LiDAR enabled technologies that facilitate a range of functions such as personnel detection through foliage (Tether, 2004), ecological measurement (Eitel *et al.*, 2016), self-driving cars (Lin *et al.*, 2020), and digital recreation of 3-dimensional (3D) objects (Raj *et al.*, 2020). However, the technological accomplishments enabled by LiDAR devices often come with a price tag beyond typical consumer budgets due to the specialized nature of LiDAR hardware components and supporting hardware/software (Queralta *et al.*, 2019). Fortunately for consumers, recent developments now provide inexpensive LiDAR hardware to consumers in a familiar portable package: cell phones. Unfortunately for organizations conscious of their security posture, this also now provides the layman with quick, inexpensive access to devices capable of nefarious activities like stealing intellectual property (Noble, 2020) and identifying burglary targets (Sturgeon, 2021).

Many collections of buildings, land, and external or internal facilities (hereby referred to as "compounds") often wish to maintain some degree of security. The actions taken to secure a compound vary, but the primary goal is to maintain some level of privacy whether that mean only allowing certain people access to the compound or simply shrouding internal facilities with external walls to prevent intelligence gathering. Despite best efforts to maintain privacy, no fool-proof method exists to prevent exploitation by potential threats due to the possibility of insider threats, espionage, etc. (Queralta *et al.*, 2019). Therefore, compound defenders must opt to implement the best possible measures to detect threats as early as possible when preventing them is not an option.

The newfound access to LiDAR-enabled cell phones increases the difficulty of a compound's defenders' ability to detect threats as early as possible. Cell phone LiDAR now enables a person to discreetly and openly carry around cell phone, as is common for people to do, and glean information about a compound from the LiDAR

sensor (Hetherington, 2021). Whether that be the width and length of the only access road to a compound or a detailed 3D external scan of a sensitive building, the security posture of a compound suffers from the intelligence that may be quickly obtained with a commercial off-the-shelf (COTS) device. Due to the fact that LiDAR emissions lie in the infrared spectrum invisible to the naked eye, defenders may worry of the requirement for costly, specialized hardware to detect foreign LiDAR emissions. Fortunately, inexpensive COTS devices also exist that aid in the detection of LiDAR.

Despite lying above the spectrum of light visible to the naked eye, some COTS security cameras capture LiDAR light emissions (Akopyan, 2016). Fortunately for defenders, this capability exists in many new security cameras as well as various security cameras manufactured in the past couple of decades, enabling defenders in some cases to leverage cameras already installed on their compounds. These cameras fail to capture the majority of unique LiDAR characteristics, such as light frequency, that differentiate it from visible light, but they do capture the high intensity and unique pattern of light produced by 3D cell phone LiDAR. In the absence of other unique characteristics, we focus on the intensity and pattern alone to identify cell phone LiDAR activity in security camera footage.

Object detection methods identify if certain objects exist within a picture according to a model trained on images with said objects accurately marked and labelled (Aker and Kalkan, 2017). In order to produce a model that accurately identifies 3D cell phone LiDAR in an image, we must first train the model on images that contain accurately marked and labelled 3D cell phone LiDAR emissions. No public collection of such data or any framework for producing such data exists to our knowledge, so this research develops a framework to collect security camera footage of 3D cell phone LiDAR emissions and produce individual frames with each LiDAR object marked and labelled for use in training an object detection machine learning model. Image thresholding algorithms allow us to isolate the LiDAR object automatically, so we investigate which algorithm produces the most accurate bounding boxes.

2. Background and related research

The rear LiDAR module on cell phones commonly emits a 2-dimensional (2D) array of beams (Figure 1) for use in applications such as 3-dimensional (3D) scanning and distance measuring (Raj *et al.*, 2020; Gallagher, 2021). This characteristic exists in multiple LiDAR-enabled cell phones, but this research uses only the rear LiDAR sensor on the iPhone 12 Pro Max (hereby referred to generically as “cell phone”). However, the research can be applied to other cell phones if their rear LiDAR sensor emits a 2D array similar to the iPhone 12 Pro Max.

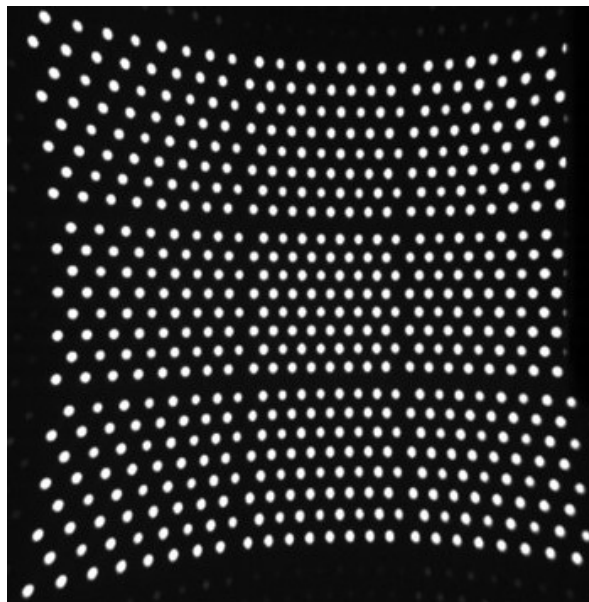


Figure 1: 2D array of LiDAR light points emitted by iPhone 12 Pro Max (Y, 2021)

Most cameras that use a complementary metal–oxide–semiconductor (CMOS) sensor can record a portion of light above the visible spectrum in the near-infrared (NIR) spectrum between 800 nanometers (nm) and 1000 nm (Vollmer, Möllmann and Shaw, 2015; Akopyan, 2016). However, physical NIR filters normally block the CMOS

sensor to prevent NIR light from distorting the image (Vollmer, Möllmann and Shaw, 2015). Cameras utilizing a CMOS sensor, such as the Lorex LNB8921BW-C used in this research, contain a Night Mode which physically moves the NIR filter out of the way of the sensor which enables the capture of NIR light (Lorex, 2021). Many LiDAR sources emit light in the NIR spectrum, enabling cameras with unfiltered CMOS sensors to capture the light (Li *et al.*, 2021).

LiDAR technologies obtain information by reflecting light off a surface and capturing the light that returns, meaning the light only contains meaningful data directly after emission from the LiDAR source and after its first reflection off a surface (Raj *et al.*, 2020). Therefore, cameras must capture LiDAR emissions either directly after they leave the LiDAR source (“direct”) or after they reflect off a single surface (“indirect”) to identify light characteristics indicative of LiDAR. Four key characteristics of LiDAR emissions help differentiate them from visible light: pulse frequency, color, intensity, and pattern (Marcoe, 2007; Raj *et al.*, 2020). The pulse frequency, color, and intensity differentiate LiDAR light from visible light in direct captures whereas the intensity and pattern differentiate between the two in indirect captures. The described capabilities of CMOS cameras and LiDAR sources indicate that both methods should be possible, but testing indicates that we can only reasonably differentiate between cell phone LiDAR emissions and visible light in indirect light captures.

With the NIR filter blocking NIR light from the CMOS sensor, our Lorex camera captures LiDAR from the Garmin LiDAR Lite 3HP as an intense white dot with purple edges that is roughly the width of a human thumb (1 inch) at a distance of 4 feet (see Figure 2). To differentiate the Garmin LiDAR from a light source in the visible spectrum, we can analyze the intensity and color components of the dot. Furthermore, we can develop a Fourier Transform model of the LiDAR emissions based on the known frequency and pulse width of the Garmin LiDAR source to compare to the Fourier Transform of the dot in the image. However, we do not know the exact frequency or pulse width of the cell phone LiDAR, and testing shows that cell phone LiDAR produces a much smaller dot at a maximum visible distance of 8 inches (see Figure 3). Therefore, analyzing the pulse frequency, intensity, and color of the cell phone LiDAR emitted directly into the camera sensor is not feasible.



Figure 2: Garmin LiDAR captured by Lorex camera with Night Mode disabled at 4 feet

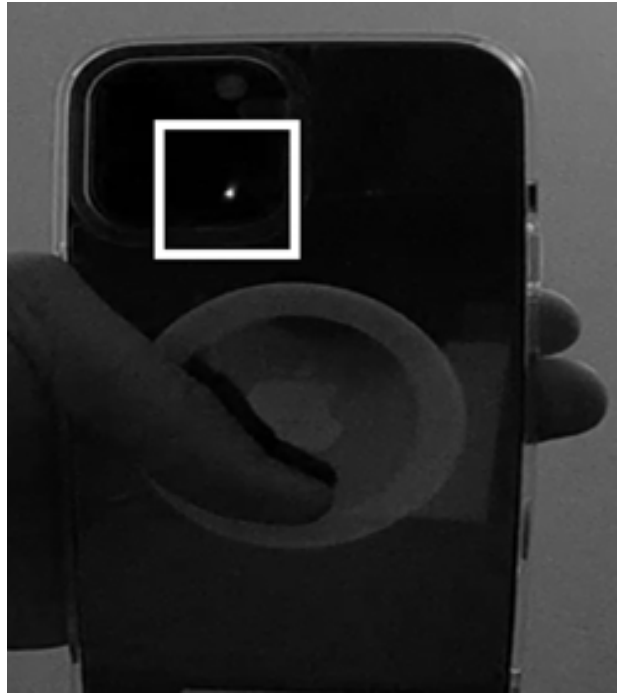


Figure 3: Cell phone LiDAR captured by Lorex camera with Night Mode disabled at 8 inches

Indirect capture provides two benefits: wider angle of collection and pattern recognition. Only a fraction of the LiDAR light reflected off a surface returns to the LiDAR sensor while the rest scatters into the environment (Raj *et al.*, 2020). This allows a secondary sensor (the camera) to capture the reflected light from a more diverse range of angles than is required in direct capture. This also allows the camera to capture all the light points in the 2D matrix emitted by the cell phone in order to identify the unique pattern (see Figure 4). Although the camera captures many of the LiDAR light points while in the presence of other light sources, isolating the light points in an image proves difficult due to the low intensity of the reflected LiDAR. Capturing the indirect LiDAR in a dark room increases the contrast of the light points, making it much easier to isolate them from the background surface. Once isolated, a machine learning model can learn the attributes of the cell phone LiDAR to identify the LiDAR in other images.

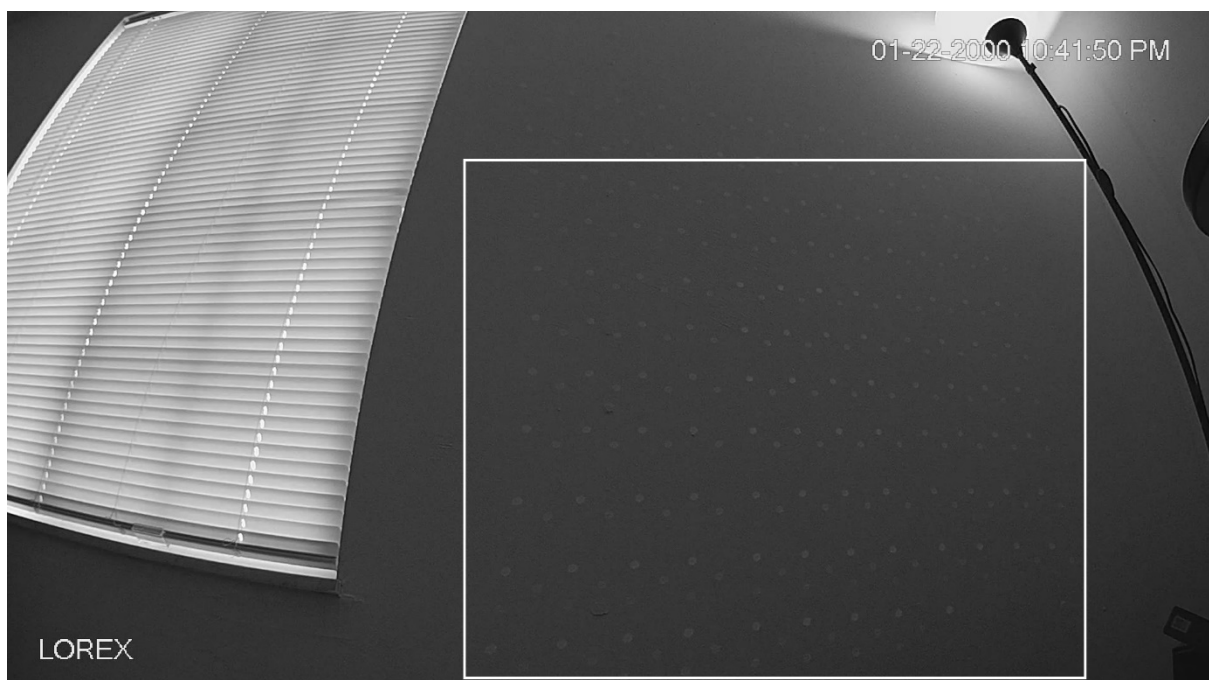


Figure 4: Indirect cell phone LiDAR emitted against a wall at 4 feet

Much research exists on both training and employing machine learning models for purposes such as human body posture analysis (Xie and Guo, 2019), common household objects identification (Noman, Stankovic and Tawfik, 2019), license plate detection on moving cars (Peker, 2019), and custom generated traffic sign detection (Kilic and Aydin, 2020). These projects all utilize the TensorFlow machine learning framework which we use in this research (TensorFlow, 2018). Other popular machine learning frameworks such as Caffe, Torch, and Theano exist, but they lack the relative GPU support, ease of use, and customization of TensorFlow (Phadnis, Mishra and Bendale, 2018). TensorFlow provides highly accessible functions thanks to its implementation in Python, a powerful high-level programming language which we use exclusively for this research (Python, 2021). The high customization of TensorFlow allows us to produce a model from a wide range of machine learning algorithms. The exact TensorFlow algorithm used falls outside of the scope of this research, but studies show that TensorFlow provides many efficient algorithms for object detection which makes it a good candidate towards which we develop training data (Phadnis, Mishra and Bendale, 2018; Kilic and Aydin, 2020).

To identify custom objects in an image, we must first train a TensorFlow machine learning model on a dataset containing said objects. Production of a training dataset employs the following generic steps: data collection and cleaning, object marking and labelling, and dataset structure generation (Kilic and Aydin, 2020). We follow this methodology but break it down into further specific steps in Section 3 to generate the training dataset.

Data collection and cleaning consists of collecting the images that contain the objects the machine learning model will learn to detect. “Cleaning” regards ensuring all images possess the same characteristics such as resolution, height and width, color mode, etc (Kilic and Aydin, 2020).

Object marking and labelling, typically the most labor-intensive part, draws a bounding box around each relevant object in an image and labels it with its object name, or class (Kilic and Aydin, 2020). A bounding box, more aptly called a minimum bounding box, describes the smallest rectangle that surrounds the entire desired object. Each bounding box contains a label of the object. A training dataset must contain accurately marked and labelled data, called “truth data”, to ensure the machine learning model learns the relevant information about the object it should detect (Kilic and Aydin, 2020). Due to the large workload required to manually mark and label many images, much research exists on methods to automatically mark and label data (Feng, Jun and Cheng, 2010; Kim, Hong and Han, 2018; Pelkmann, Tharwat and Schenck, 2020; Roh, Heo and Whang, 2021). In the absence of automatic methods, tools such as Labellmg allow us to mark and label each image by hand (tzutalin, 2021).

Calculating the bounding boxes relies on the concept of image thresholding to detect the edges of the light points. Image thresholding algorithms segment images based on detected object edges using various analysis techniques. Two popular algorithms perform such analysis efficiently: Otsu’s algorithm and Yen’s algorithm. Otsu’s image thresholding algorithm minimizes the intra-class variance of pixel color values in a grayscale image to dynamically calculate a threshold value (Otsu, 1979). Yen’s algorithm defines a cost function which it minimizes to calculate a threshold value (Yen, Chang and Chang, 1995). When applying the threshold value as a binary filter on a grayscale image, all pixels with a color value below the threshold turn black (grayscale value 0), while all those equal to or above the threshold turn white (grayscale value 255). Only the high intensity pixels remain, allowing for the trivial calculation of the bounding box from the minimum and maximum X and Y coordinates of all non-black pixels.

The accuracy of a calculated bounding box regards its proximity to the true bounding of the object. Many machine learning applications gauge accuracy with a metric called Intersection over Union (IoU) which compares the intersecting area of two bounding boxes to the total area covered by both boxes. A perfect calculated bounding box has an IoU of 1 (completely intersecting) while a completely inaccurate calculated bounding box has an IoU of 0 (not intersecting at all) (Rezatofighi *et al.*, 2019).

3. Project design

We developed a framework to collect and parse video footage that consists of three distinct parts: the cell phone, the security camera, and the parsing software.

As shown in Figure 5, we place the cell phone and security camera in a dark room approximately 5 feet deep and 4 feet wide. Both the cell phone and camera stand at the same height 4.5 feet off the ground pointed directly horizontal at the back wall. The camera points from the left wall approximately 3 feet from the back wall, while

the phone points from the front wall approximately 5 feet from the back wall. The field of view of both devices intersect, enabling the camera to record the indirect LiDAR emissions from the cell phone.

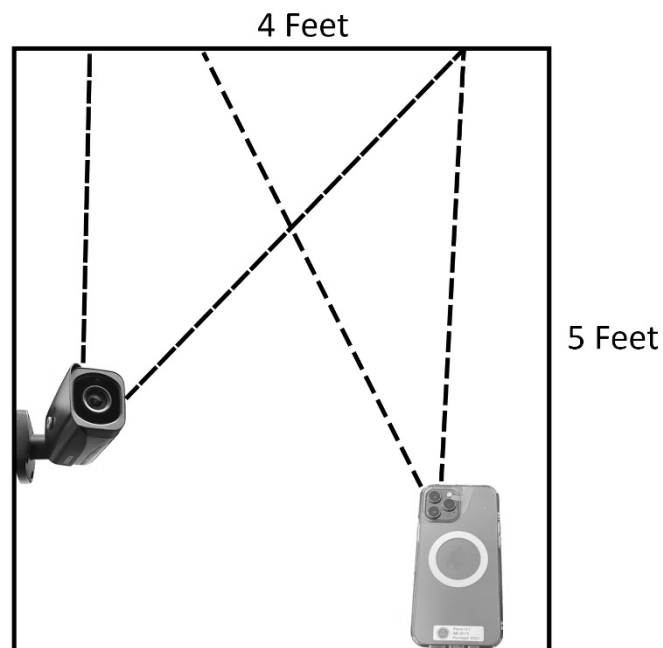


Figure 5: Cell phone and camera set up in dark room

3.1 iPhone 12 pro max

We use the iPhone 12 Pro Max to emit LiDAR light from the rear LiDAR sensor at a wall which reflects off for collection by the security camera. Apple provides a sample project that enables distance data collection with the rear LiDAR module (Apple, 2020). We slightly modify this project for ease of use, but the project as provided by Apple sufficiently emits the 2D LiDAR matrix (Figure 1) for collection by the security camera. We use the software XCode 12 on a 2020 M1 MacBook Pro to deploy the project as an app for the cell phone (Apple, 2021). Once deployed to the phone, we simply open the app and point it at the wall to constantly emit LiDAR light.

3.2 CMOS security camera

We use the Lorex LNB8921-C IP security camera to record video footage of the indirect LiDAR emissions. Preliminary testing shows successful collection of indirect LiDAR emission from the cell phone when we enable Night Mode on the camera. Various shutter speeds accurately record the LiDAR emissions, but testing shows that a shutter speed of 1/30 performs best at collecting video footage to process for training a future machine learning model. We record the video footage with a resolution of 704x480 at a frame rate of 15 frames per second (FPS). All of the aforementioned settings must be set manually in the camera's local web interface before recording footage with the Python script detailed in the following section.

3.3 Parsing software

We implemented the parsing software as a Python 3 script that primarily utilizes two libraries, TensorFlow 2.0 and OpenCV, which henceforth is referred to generically as Python, TensorFlow, and OpenCV respectively. OpenCV provides a wide range of real-time computer vision functions that we leverage for the majority of the image processing (OpenCV, 2021).

The bulk of this research focused on developing Python code to parse the video footage from the Lorex security camera and label the indirect LiDAR object data in each frame according to the standards of the machine learning training requirements for TensorFlow. This results in a dataset containing:

1. Each frame saved as an image
2. Comma Separated Values (CSV) file containing a bounding box and object label for the indirect LiDAR object in each frame
3. Class file containing LiDAR object label
4. Record files for training and evaluating machine learning model

This dataset contains everything needed for training an object identification machine learning model with TensorFlow. The following section details the exact methodology of how we implement the parsing software to generate the dataset.

4. Research methodology

To produce the dataset for training a machine learning model, our Python script collects and parses video footage in the following steps:

1. Record video footage from the security camera
2. Process the footage to generate bounding boxes and object label for indirect LiDAR emissions in each video frame
3. Produce data files required to train machine learning model

We perform step 2 twice on the same footage from step 1: once using Otsu's algorithm and once using Yen's algorithm. We use Labelling to manually mark and label the bounding boxes for each frame in the video from step 1 which we use as truth data to analyse the performance of each image thresholding image.

4.1 Record video footage

The Lorex security camera readily exports footage via the Real Time Streaming Protocol (RTSP), a standardized protocol for controlling delivery of data with real-time properties such as live video (Schulzrinne, Rao and Lanphier, 1998). This allows OpenCV to interface directly with the camera and record video footage to a computer for processing. With the camera settings configured according to the Background section, the FPS of 15 must also be provided to OpenCV when writing the video file to disk to ensure it matches the frame rate at which the camera recorded it.

4.2 Process the video footage

With the video containing indirect LiDAR emissions recorded to disk, we then process the video footage into the training dataset as follows:

1. Separate the video footage into frames
2. Calculate the threshold value for an image mask to isolate the LiDAR emission points
3. Calculate the bounding box for the light visible in the masked image
4. Write data for training use

4.2.1 Separate the video footage into frames

When reading the video file from disk, we must provide the video height, width, FPS, and encoding format to OpenCV according to the Lorex security camera settings. The Python code splits one video of indefinite length into individual frames, represented as OpenCV images, for which the next step calculates image mask thresholds.

4.2.2 Calculate threshold for image mask

For later use in calculating the bounding box of the LiDAR object in each frame, we use Otsu's and Yen's image thresholding algorithms in separate executions to isolate the LiDAR object from the original image (Figure 6(a)). This produces the image mask seen in 6(b) which we use to calculate the bounding box.

4.2.3 Calculate bounding box

The ideal bounding box encapsulates every white pixel in the image mask without encapsulating any black pixel not required to encapsulate a white pixel. Although this can be manually calculated from the minimum and maximum X and Y coordinates of all white pixels, OpenCV provides much more efficient functions that find the minimum bounding box for the contours (object edges) found in an image. Since we effectively provide the contours in the image mask, OpenCV quickly identifies them and calculates the bounding box for the LiDAR object in the image as seen in 6(c).

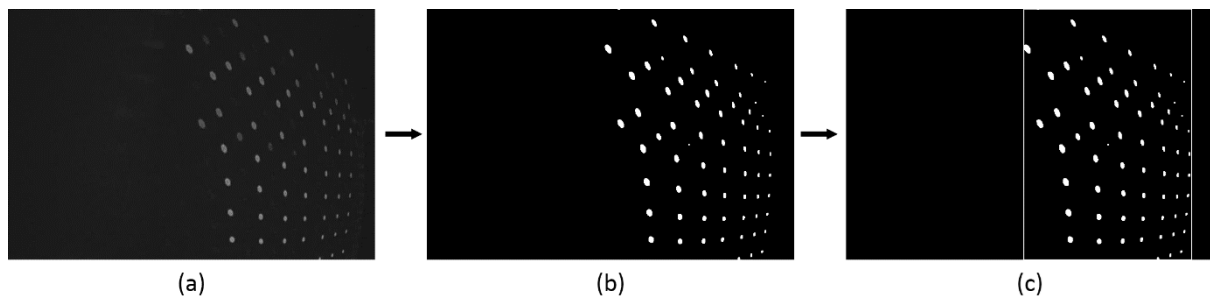


Figure 6: Image thresholding is applied to the original image (a) to derive the image mask (b) and calculate the bounding box (c)

4.2.4 Write data

With the bounding box calculated and object label known, we write the data files described in Section 3.3 for later use in training a machine learning model. We use OpenCV to save the individual frames obtained in Section 4.2.1 as PNG files. The CSV file, class file, and record files must meet specific format requirements, so we use a publicly available Python script to generate them (Meneghetti, 2020). This script formats our single “LiDAR” object label into the proper class file, as well as aggregates each image’s file name, width, height, object label, and bounding box parameters into a CSV file which it then combines with the class file to generate the record file.

5. Results and analysis

We calculate the Intersection over Union (IoU) for the bounding boxes resulting from both Otsu’s and Yen’s image thresholding algorithms as compared to our manually calculated ground truth bounding boxes. The comparison yields 150 IoU measurements, one for each image. The Wilcoxon signed rank test comparing the 150 IoU values from both Otsu’s and Yen’s algorithm reports a P-value $1.05289e-43$ indicating that Otsu’s algorithm undoubtedly performs better than Yen’s algorithm. Considering the small relative IoU standard deviations for both algorithms, a similar result can be seen intuitively in Table 7. The IoU mean of Otsu’s algorithm lies more than 10 standard deviations above the IoU mean of Yen’s algorithm, clearly indicating better accuracy. Both algorithms appear to produce highly accurate bounding boxes based on visual analysis, but this statistical analysis proves Otsu’s algorithm is the best choice for future use in producing data for training a machine learning model.

Table 7: Intersection over Union results for both image thresholding algorithms

Algorithm	IoU Mean	IoU Standard Deviation
Otsu’s	0.995169	0.0162411
Yen’s	0.926803	0.0556291

6. Conclusion

The resulting framework allows us to easily record security camera footage that contains indirect cell phone LiDAR emissions and automatically mark and label the LiDAR object in each frame, saving a large amount of time that would otherwise be spent manually producing the same results. The choice of image thresholding algorithm directly influences the accuracy of the bounding boxes which behooves us to choose the best algorithm possible to produce accurate training data for a machine learning model. Testing shows that Otsu’s image thresholding algorithm produces training data with 99.5% accurate bounding boxes, making it a better choice than Yen’s algorithm. Therefore, we plan to utilize Otsu’s image thresholding algorithm in the future to produce a large number of labelled images with which we can train a machine learning model that accurately detects LiDAR objects in a wide range of images.

Acknowledgements

We thank the government employees Ashleigh Coker and Capt Frankie Cruz (AFRL/Ryaa) for their aid in hardware acquisition.

References

Aker, C. and Kalkan, S. (2017) ‘Using deep networks for drone detection’, *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017*. doi: 10.1109/AVSS.2017.8078539.

- Akopyan, V. (2016) *A modified DSLR does not make a good multispectral sensor, Quickbird*. Available at: <https://blog.quickbird.uk/why-a-modified-dslr-does-not-make-a-good-multispectral-sensor-6766a9270820> (Accessed: 8 June 2021).
- Apple (2020) *Displaying a Point Cloud Using Scene Depth*. Available at: https://developer.apple.com/documentation/arkit/environmental_analysis/displaying_a_point_cloud_using_scene_depth (Accessed: 10 September 2021).
- Apple (2021) *Xcode 13 Overview*. Available at: <https://developer.apple.com/xcode/> (Accessed: 10 September 2021).
- Eitel, J. U. H. et al. (2016) 'Beyond 3-D: The new spectrum of lidar applications for earth and ecological sciences', *Remote Sensing of Environment*. Elsevier Inc., pp. 372–392. doi: 10.1016/j.rse.2016.08.018.
- Feng, Q., Jun, H. and Cheng, Z. K. (2010) 'An algorithm for multi-label learning based on minimum threshold', *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, pp. 2793–2797. doi: 10.1109/WCICA.2010.5554938.
- Gallagher, W. (2021) *How to use the LiDAR scanner in iPhone 12 Pro, Apple Insider*. Available at: <https://appleinsider.com/articles/21/03/02/how-to-use-the-lidar-scanner-in-iphone-12-pro> (Accessed: 7 September 2021).
- Hetherington, E. (2021) *LiDAR on a phone - a new way to capture 3D data - Resource Centre, Esri UK & Ireland*. Available at: <https://resource.esriuk.com/blog/lidar-on-a-phone-a-new-way-to-capture-3d-data/> (Accessed: 14 September 2021).
- Kilic, I. and Aydin, G. (2020) 'Traffic Sign Detection and Recognition Using TensorFlow's Object Detection API with A New Benchmark Dataset', *2020 International Conference on Electrical Engineering, ICEE 2020*. doi: 10.1109/ICEE49691.2020.9249914.
- Kim, K., Hong, Y. G. and Han, Y. H. (2018) 'General labelled data generator framework for network machine learning', *International Conference on Advanced Communication Technology, ICACT*, 2018-Febru, pp. 127–131. doi: 10.23919/ICACT.2018.8323670.
- Li, N. et al. (2021) 'Spectral imaging and spectral LIDAR systems: moving toward compact nanophotonics-based sensing', *Nanophotonics*, 10(5), pp. 1437–1467. doi: 10.1515/NANOPH-2020-0625.
- Lin, G. H. et al. (2020) 'Self-driving Deep Learning System based on Depth Image Based Rendering and LiDAR Point Cloud', *2020 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-Taiwan 2020*. doi: 10.1109/ICCE-TAIWAN49838.2020.9258010.
- Lorex (2021) *4K Ultra HD Resolution 8MP Outdoor IP Camera, 200ft Night Vision*. Available at: <https://www.lorextechnology.com/4k-security-camera/4k-ultra-hd-8mp-nocturnal-ip-camera/LNB8921BW-1-p> (Accessed: 10 September 2021).
- Marcoe, K. (2007) *LIDAR an Introduction and Overview*. Available at: http://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Marcoe_LiDAR.pdf (Accessed: 13 June 2021).
- McManamon, P. F., Kamerman, G. and Huffaker, M. (2010) 'A history of laser radar in the United States', in *Laser Radar Technology and Applications XV*. SPIE, p. 76840T. doi: 10.1117/12.862562.
- Meneghetti, D. D. R. (2020) *detection_util_scripts/generate_tfrecord.py, GitHub*. Available at: https://github.com/douglasrizzo/detection_util_scripts/blob/master/generate_tfrecord.py (Accessed: 8 September 2021).
- Molebny, V., Kamerman, G. and Steinvall, O. (2010) 'Laser radar: from early history to new trends', in *Electro-Optical Remote Sensing, Photonic Technologies, and Applications IV*. SPIE, p. 783502. doi: 10.1117/12.867906.
- Noble, J. (2020) *FIA bans 3D cameras as part of F1 car copying clampdown, Autosport*. Available at: <https://www.autosport.com/f1/news/fia-bans-3d-cameras-as-part-of-f1-car-copying-clampdown-4977570/4977570/> (Accessed: 15 September 2021).
- Noman, M., Stankovic, V. and Tawfik, A. (2019) 'Object Detection Techniques: Overview and Performance Comparison', *2019 IEEE 19th International Symposium on Signal Processing and Information Technology, ISSPIT 2019*. doi: 10.1109/ISSPIT47144.2019.9001879.
- OpenCV (2021) *Home - OpenCV*. Available at: <https://opencv.org/> (Accessed: 10 September 2021).
- Otsu, N. (1979) 'A Threshold Selection Method from Gray-Level Histograms', *IEEE Trans Syst Man Cybern*, SMC-9(1), pp. 62–66. doi: 10.1109/TSMC.1979.4310076.
- Peker, M. (2019) 'Comparison of Tensorflow Object Detection Networks for Licence Plate Localization', *Proceedings - 2019 IEEE 1st Global Power, Energy and Communication Conference, GPECOM 2019*, pp. 101–105. doi: 10.1109/GPECOM.2019.8778602.
- Pelkmann, D., Tharwat, A. and Schenck, W. (2020) 'How to Label? Combining Experts' Knowledge for German Text Classification', *Proceedings - 2020 7th Swiss Conference on Data Science, SDS 2020*, pp. 61–62. doi: 10.1109/SDS49233.2020.00023.
- Phadnis, R., Mishra, J. and Bendale, S. (2018) 'Objects Talk - Object Detection and Pattern Tracking Using TensorFlow', *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICT 2018*, pp. 1216–1219. doi: 10.1109/ICICT.2018.8473331.
- Python (2021) *Welcome to Python.org*. Available at: <https://www.python.org/> (Accessed: 8 September 2021).
- Queralta, J. P. et al. (2019) 'FPGA-based Architecture for a Low-Cost 3D Lidar Design and Implementation from Multiple Rotating 2D Lidars with ROS', *Proceedings of IEEE Sensors*, 2019-October. doi: 10.1109/SENSORS43011.2019.8956928.
- Raj, T. et al. (2020) 'A Survey on LiDAR Scanning Mechanisms', *Electronics (Switzerland)*. MDPI AG, p. 741. doi: 10.3390/electronics9050741.

- Rezatofighi, H. *et al.* (2019) 'Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, pp. 658–666. Available at: <https://arxiv.org/abs/1902.09630v2> (Accessed: 15 September 2021).
- Roh, Y., Heo, G. and Whang, S. E. (2021) 'A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective', *IEEE Transactions on Knowledge and Data Engineering*, 33(4), pp. 1328–1347. doi: 10.1109/TKDE.2019.2946162.
- Schulzrinne, H., Rao, A. and Lanphier, R. (1998) *Real Time Streaming Protocol (RTSP)*, *Internet Engineering Task Force*. Available at: <https://datatracker.ietf.org/doc/html/rfc2326> (Accessed: 10 September 2021).
- Sturgeon, R. (2021) *The Privacy and Security Threat of the iPhone 12 Pro LIDAR Sensor, Mac O'Clock*. Available at: <https://medium.com/macoclock/the-privacy-and-security-threat-of-the-iphone-12-pro-lidar-sensor-bb6de2bc4dd9> (Accessed: 15 September 2021).
- TensorFlow (2018) *Basics of Machine Learning*, TensorFlow. doi: 10.1007/978-3-319-77800-6_8.
- Tether, T. (2004) 'Subcommittee on Terrorism, Unconventional Threats and Capabilities, House Armed Services Committee, U.S. House of Representatives', 25 March. Available at: [https://www.darpa.mil/attachments/TestimonyArchived\(March 25 2004\).pdf](https://www.darpa.mil/attachments/TestimonyArchived(March%2025%202004).pdf) (Accessed: 14 June 2021).
- tzutalin (2021) *LabelImg*, Github. Available at: <https://github.com/tzutalin/labelImg> (Accessed: 15 September 2021).
- Vollmer, M., Möllmann, K.-P. and Shaw, J. A. (2015) 'The optics and physics of near infrared imaging', *Education and Training in Optics and Photonics: ETOP 2015*, 9793(8), p. 97930Z. doi: 10.1117/12.2223094.
- Xie, L. and Guo, X. (2019) 'Object detection and analysis of human body postures based on TensorFlow', *Proceedings - 2019 IEEE International Conference on Smart Internet of Things, SmartIoT 2019*, pp. 397–401. doi: 10.1109/SMARTIOT.2019.00070.
- Y, M. (2021) *Apple LIDAR Demystified: SPAD, VCSEL, and Fusion....*. Available at: <https://3da.medium.com/apple-lidar-demystified-spad-vcsel-and-fusion-aa9c3519d4cb> (Accessed: 8 June 2021).
- Yen, J. C., Chang, F. J. and Chang, S. (1995) 'A New Criterion for Automatic Multilevel Thresholding', *IEEE Transactions on Image Processing*, 4(3), pp. 370–378. doi: 10.1109/83.366472.

Securing InfiniBand Networks with the Bluefield-2 Data Processing Unit

Noah Diamond¹, Scott Graham¹ and Gilbert Clark²

¹Air Force Institute of Technology, Wright-Patterson AFB, USA

²Sensors Directorate, Air Force Research Laboratory, Wright-Patterson AFB, USA

Noah.Diamond@afit.edu

Scott.Graham@afit.edu

Gilbert.Clark.2@us.af.mil

Abstract: Interest in securing InfiniBand networks with encryption is growing. However, the performance benefit realized by InfiniBand's use of Direct Memory Access (DMA) to bypass the kernel and avoid intervention from host Central Processing Units (CPUs) is at odds with IP datagram encryption techniques. Encryption forces data through the CPU before transmission and decryption, incurring multiple clock cycles. The Bluefield-2 Data Processing Unit (DPU) is Nvidia-Mellanox's latest system on chip that combines a high-performance, programmable processor, network interface card (NIC), and flexible hardware accelerators. This research characterizes the Bluefield-2's capability to accelerate IPsec encryption in hardware. Results show that the Bluefield-2's hardware accelerators are capable of supporting a secure IPsec tunnel with a throughput of nearly 16 Gb/s. Offloading IPsec encryption operations to the hardware accelerators on the Bluefield-2 is a promising method for adding confidentiality, integrity, and authentication to InfiniBand networks.

Keywords: Cybersecurity, InfiniBand, Bluefield-2 DPU, Hardware Acceleration, IPsec

1. Introduction

InfiniBand is an industry leading high-bandwidth, low-latency interconnect for hyperscale datacenters and high-performance computing (HPC) clusters. Previous research identified that native InfiniBand clear text key exchange makes the IBA vulnerable to man-in-the-middle (MiTM), denial of service (DoS), and replay attacks (Lee and Kim, 2007). These vulnerabilities have mostly been ignored during InfiniBand's development because datacenters and HPCs are assumed to be physically secure. However, interest in securing InfiniBand with encryption is growing. While methods of encryption like IP security protocol (IPsec) can increase data confidentiality, integrity, and authentication, encryption is inherently computationally demanding and increases the compute load placed on central processing units (CPUs). Data processing units (DPUs) are a new class of programmable processor designed to assist CPUs and meet the tremendous computational demand found in hyperscale datacenters and HPCs. DPUs are a system on chip (SoC) that combines a high-performance, programmable processor, network interface card (NIC), and flexible hardware accelerators. This additional hardware is capable of supporting network management and security applications while requiring little interaction from host CPUs. They are specifically designed for processing datacenter software for virtualization, networking, storage, and security. Although previous research suggested that DPU processors are capable of line-rate encryption, Nvidia Mellanox's Inova Flex SmartNIC and Bluefield-1 DPU were not inherently able to accelerate offloaded IPsec encryption operations in hardware. This research seeks to characterize the capability of Nvidia Mellanox's Bluefield-2 DPU's hardware accelerators to encrypt Ethernet traffic.

2. Background

2.1 InfiniBand Architecture

The InfiniBand Architecture (IBA) is a low-latency, high-bandwidth fabric architecture for server input/output (I/O) and inter-server communication. It was developed by the InfiniBand Trade Association (IBTA) to meet the growing demand for a high-performance interconnect capable of providing low-latency and high-bandwidth in hyperscale data centers and HPCs. InfiniBand provides a standard of reliability, availability, performance, and scalability that far surpasses that of bus-oriented I/O architectures (Pfister, 2001). InfiniBand leads the industry in both chassis backplane applications as well as in external copper and fiber optic connections (*InfiniBand Trade Association*, 2021). Six out of the top ten supercomputers in the world use IBA as their core interconnect and it is currently in use in thousands of datacenters, HPC clusters, and embedded applications. (Strohmaier *et al.*, 2021).

Widespread demand for high-performance, scalable, and reliable networks in a diverse set of applications has promoted interest in InfiniBand networks. Amidst the rapid development of the IBA, developers have paid more

attention to performance and cost efficiency than to security. Lee and Kim (2007) and Rothenberger et al. (2021) state that there are numerous security loopholes within the IBA that have been revealed and, consequently, the design of secure clusters has recently surfaced as a critical issue.

2.2 Bluefield-2 DPU

The Nvidia-Mellanox's Bluefield-2 DPU combines a ConnectX-6 DX network adapter with an array of Advanced RISC Machine (ARM) cores and hardware accelerators. The Bluefield-2 operates as an independent system that communicates with its host over 16 lanes of fourth generation Peripheral Component Interconnect express (PCIe), offering a theoretical transfer rate of 256 Gbps. Alternatively, the card may be used in systems with third generation PCIe, but the maximum theoretical transfer rate is substantially lower at only 128 Gbps. The card itself includes two multi-function 100 Gbps ports, 16 GB of local Double Data Rate 4 Random-Access Memory (DDR4 RAM), 8 ARM Cortex A72 processors, and local persistent storage. The transfer rate of the Bluefield-2's DDR4 RAM is 3200 T/s. The card runs a tailored version of Linux provided by NVIDIA Mellanox allowing developers to both develop new applications and deploy existing applications directly onto the card itself. These applications can process and modify traffic before it is ever seen on the host. The Bluefield-2s, therefore, can host a wide variety of applications and services for networking, storage, and security ('NVIDIA BLUEFIELD-2 DPU DATA CENTER INFRASTRUCTURE ON A CHIP', 2021).

Even though these desirable properties have been established, the high data throughput supported by the card can quickly overwhelm its available processors and memory should all traffic be directed through Linux itself. To address this, the card offers several different hardware offload and acceleration features that can operate directly on network traffic without routine involvement from the ARM CPU. This allows the ARM multi-core CPU to orchestrate the hardware to perform operations on traffic at high rate instead of rather than trying to process all the traffic directly.

2.3 IPsec

This paper focuses on a unique IPsec hardware acceleration feature on the Bluefield-2. IPsec is a protocol that is commonly used to provide security at the network layer. This research uses Encapsulation Security Payload (ESP) protocol to provide confidentiality, integrity, and authentication at the network layer (Kurose and Ross, 2017).

2.3.1 IPsec Datagram

Figure 1. shows the format of an IPsec datagram using ESP and tunnel mode. The IPsec datagram still meets the requirements of an IPv4 datagram. The IPsec datagram's payload consists of an ESP header, the original IP datagram, an ESP trailer, and an authentication field.

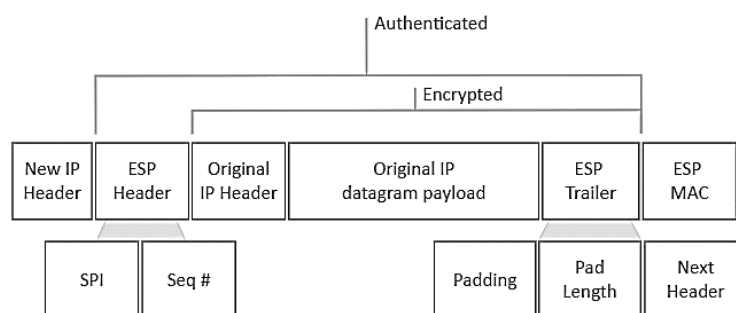


Figure 1: IPsec datagram format (Kurose and Ross, 2017)

These headers and trailers create additional overhead and must be accounted for when configuring the Maximum Transmission Unit (MTU) of network interfaces. In total, the protocol suite can add over 100 bytes of overhead to IP datagrams. As a result, care must be taken to ensure that the payload, when combined with the IPsec headers, does not exceed the MTU of the network link. If it does, the resulting packets could be fragmented or dropped. Further, to optimize operation, the maximum segment size (MSS) of any TCP streams traveling over an IPsec tunnel should also account for the reduced MTU.

2.4 Tools

2.4.1 iPerf3

iPerf3 actively measures the maximum achievable bandwidth of IP datagram networks. iPerf3 provides various tuning options and allows parameters like transport protocol, MSS, and multi-threading to be selected during bandwidth tests. An iPerf3 server and client are run on the workstations during this research to record real throughput measurements.

2.4.2 Libreswan

Libreswan is a free software implementation of IPsec that supports customized configurations. This research uses symmetric key encryption. Keys are defined by the user and located in Libreswan IPsec tunnel configuration files.

2.4.3 Top

Top is a Linux command line tool used to show real-time summary of the system which includes CPU utilization and Luniux processes. The real-time statistics provided by top are useful for verifying that the Bluefield-2s are the bottleneck in the network rather than the host CPUs generating traffic.

2.4.4 tcpdump

Tcpdump is a free command-line packet analyzer. Tcpdump is useful for monitoring network traffic, and is used in this research to verify the network is properly configured. For example, this research studies the effect encryption has on network performance. Tcpdump displays the type of packet an interface is receiving and an ESP header indicates that the traffic is encrypted.

2.4.5 R

R is a free statistical computing and graphics software. The figures and statistical analysis in this research were generated using R.

2.5 Related work

2.5.1 A Comprehensive Framework for Enhancing Security in InfiniBand Architecture

Prior research investigated security vulnerabilities within the IBA and proposed a comprehensive security framework for securing the IBA. Lee and Kim identified that most of the vulnerabilities found in InfiniBand networks stem from InfiniBand's plain text key management scheme. In response, they proposed light weight encrypted key management schemes that add confidentiality, authentication, and availability to InfiniBand networks. For confidentiality and authentication, Lee and Kim created a partition-level and queue-pair (QP) level secret key management scheme. Further, they demonstrated how IBA accommodates GCM with minor modifications to the IBA specification. GCM combines encryption and authentication by using Galois Field (GF) multiplication and counter (CTR) mode encryption. For availability, Lee and Kim proposed a stateful ingress filtering (SIF) system which is enabled only when there is a DoS attack using invalid IBA keys.

Lee and Kim ran a variety of simulations to test their hardware-based encryption approach against a variety of attacks. The results showed marginal performance degradation due to the adoption of encryption and authentication algorithms. Future research could use the Bluefield-2 to implement Lee and Kim's secure key management scheme.

2.5.2 AFIT: Securing InfiniBand

Schmitt was the first master's student at the Air Force Institute of Technology (AFIT) to study InfiniBand. His cyber vulnerability assessment of InfiniBand Networks paved the way for subsequent research investigating the capabilities of several Nvidia-Mellanox channel adapters. The following year, Mireles researched the possibility of encrypting at line rate (e.g., >100Gb/s) using hardware acceleration technology using the Nvidia-Mellanox Innova IPsec Adapter and the Innova 2 Flex Adapter. He concluded that both adapters were unable to hardware accelerate IPsec encryption given the most recent firmware and FPGA images. Lastly, Hintze investigated the capability of the Bluefield-1 DPU to hardware accelerate IPsec encryption. He found that the Bluefield-1 was unable to hardware accelerate IPsec encryption. This research builds off previous research at AFIT by investigating the Bluefield-2 capability to hardware accelerate IPsec.

2.5.3 Performance characteristics of the Bluefield-2 DPU

Liu et al. focused on characterizing the networking and computing aspects of the Bluefield-2 in a 100 Gb/s Ethernet Environment. They ran a wide variety of benchmarks on the Bluefield-2 and compared the results to the performance of several, commercially available host CPUs. Their research complements this research by concluding that the offloading encryption/decryption is a promising feature of the Bluefield-2. They identified that offloading encryption/decryption to the Bluefield-2 could significantly save CPU cycles for applications running on the host and reduce function latency. This research further investigates the benefit of offloading encryption/decryption to the Bluefield-2.

3. Experimental Setup

3.1 Network Topology

The network topology used for this research is shown in Figure 2. The Bluefield-2s on each workstation ride on sixteen lanes of PCIe gen 3 and are connected in tandem by a 100 Gb/s fiber optic cable. Figure 2. also illustrates how traffic moves through each Bluefield-2 while operating in SmartNIC mode (also known as Embedded Mode). While in SmartNIC mode, the Arm subsystem takes control of the Bluefield-2, and a virtual bridge, Open v Switch (OVS), forwards traffic to the appropriate interface on the SmartNIC. The Bluefield-2s use OVS in conjunction with OpenFlow (OF) rules to configure hardware offload and other filtering functions.

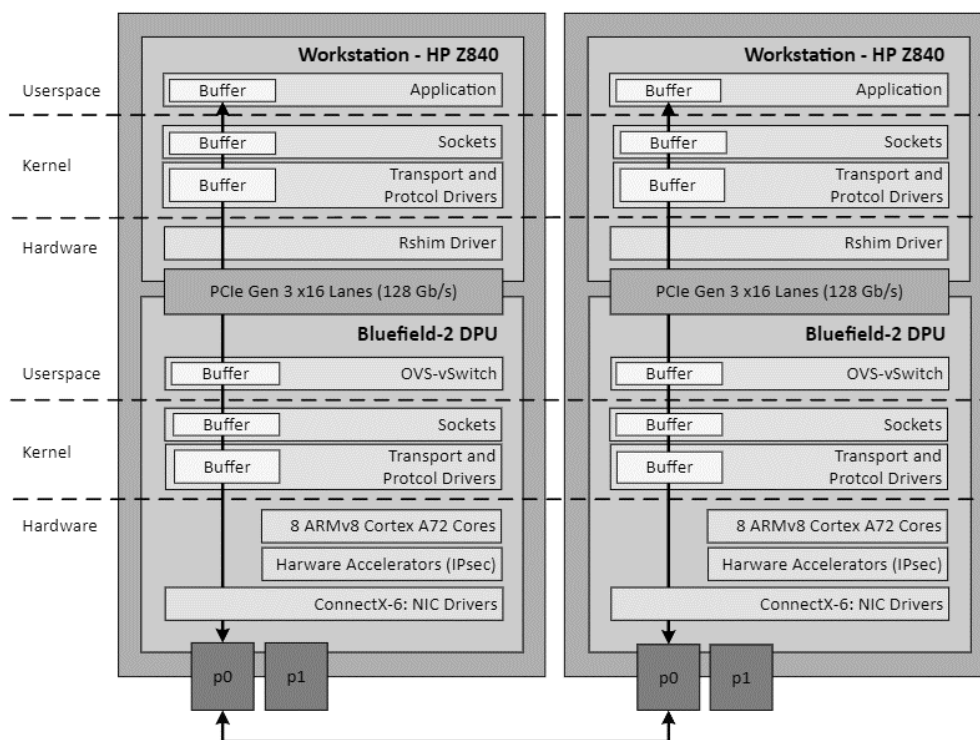


Figure 2: Network Topology: No hardware acceleration

3.1.1 Workstations

Bluefield-2s are installed on two identical HP Z840 workstations. The HP Z840s have up to PCIe gen 3 which are capable of 128 GT/s using sixteen lanes. The PCIe gen 3 should be able provide sufficient throughput for this research even though the Bluefield-2s are compatible with PCIe gen 4. Additionally, the HP Z840s have 20 Intel Xeon Cores, 256 GB of RAM, and a 1 TB hard drive.

4. Methodology

This research focuses on characterizing the end-end throughput achievable using two the Bluefield-2 DPU's hardware accelerators while using Ethernet and TCP at the link and transport layers respectively. Four application scenarios are required to fully characterize the Bluefield-2 hardware accelerators because encryption and hardware acceleration are two level factors (2^2). In combination, the four application scenarios

form a full factorial design. Full factorial experimental designs minimize aliasing and aid in reducing noise from the experimental results. Each experiment in this research was randomized in an attempt to further reduce noise incurred by run order.

4.1 Variables

The control variables are those held constant over the execution of the various scenarios and experiments of this research. These are listed in Table 1.

Table 1: Control Variables

Variable	Value	Description
Test Duration	10 seconds	iPerf3 takes active bandwidth measurements for each second for the duration of the test

Independent variables are specific to the scenarios aimed at evaluating and optimizing the performance of the network. These are listed in Table 2.

Table 2: Independent Variables

Variable	Low Level	High Level	Units	Description
MSS	890	8900	Bytes	Maximum Segment Size
Thread	1	8	N/A	iPerf3 test threads
Performance Setting	ONDEMAND	PERFORMANCE	N/A	Performance setting of host CPU
Direction	WS3 → WS4	WS3 → WS4	N/A	Direction of iPerf3 Test

Response variables are correlated to the performance of the network given a specific scenario. These are listed in Table 3.

Table 3: Response Variables

Variable	Units	Description
Throughput	Gb/s	Maximum rate of data transfer across a given path

4.2 Application Scenarios

This research tests the effect two factors have on throughput in the network described in Section 3.1. First, this research captures the performance degradation IPsec has on network performance under normal conditions. Second, this research measures the performance benefit hardware offloading provides because processing network traffic in software is often slower than hardware implementations.

4.2.1 Sending plain text without hardware acceleration

The first application scenario of this research establishes the baseline performance of the Bluefield-2s using Transmission Control Protocol (TCP) and Ethernet at the transport and link layers respectively. Traffic at each Bluefield-2 is inspected by a virtual bridge.

4.2.2 Sending plain text with hardware acceleration

This scenario offloads *all* traffic through the hardware accelerators onboard the Bluefield-2 using OpenFlow rules. Sending all traffic to the hardware maximizes the network performance because network traffic has little to no interaction with software at the Bluefield-2.

4.2.3 Sending encrypted (IPsec) without hardware acceleration

This scenario implements IPsec using the open source LibreSwan library configured in tunnel mode. IP datagrams are encrypted by the workstation CPUs in this scenario, and the virtual switch on each Bluefield-2 forwards encrypted datagrams from the host across the network.

4.2.4 Sending encrypted (IPsec) with hardware acceleration

Lastly, this scenario implements IPsec using the Bluefield-2 hardware accelerators. Each Bluefield-2 is configured in legacy single root I/O virtualization (SR-I/OV) mode and has full hardware offload enabled. Further, IPsec full offload rules specific to the network topology shown in Section 3.1 are loaded on the Bluefield-2. Each Bluefield has a matching set of offload rules and keys so that an IPsec tunnel is properly configured between the two Bluefield-2s. IPsec keys are manually distributed in this research.

4.3 Assumptions

Nvidia-Mellanox provides a tuning tool for the Bluefield-2 optimizes network performance. This research does not use the tuning tool because reverting the system to the original state would require that the operating systems and other software to be reinstalled on each workstation and Bluefield. This research assumes that the tuning tool would improve system performance, but the results presented by this research are still representative of the impact encryption and hardware offload have on system performance.

4.4 Data Collection and Verification

4.4.1 Full Factorial Design

Confounding variables and uncontrolled factors can introduce noise into experiment results. This research applies the Kruskal-Wallis analysis of variance test on a full factorial design to identify factors that have a significant effect on the response variable, throughput. Table 4 illustrates how factor levels are set during each trial of the full factorial design. The full factorial design applies to all four of the application scenarios described in Section 4.2.

Table 4: Full Factorial Design

Treatment	MSS	Thread	Performance	Direction
1	-	-	-	-
2	+	-	-	-
3	-	+	-	-
	⋮	⋮	⋮	
15	-	+	+	+
16	+	+	+	+

4.4.2 Performance Characterization

Factors introducing noise into the full factorial design are controlled in this experiment. This is accomplished by holding them constant at the level that produces the greatest average throughput in the full factorial design. Noise is further reduced by randomization of trials. The Kruskal-Wallis Test assumes that experimental results are independent of run order and randomizing can help disperse noise caused by the run order of an experiment. This experiment characterizes the performance of the Bluefield-s2 DPU according to the four application scenarios described in Section 4.2.

4.4.3 Verification

Hosts on either end of the network topology described in Section 3.1, see network traffic in plain text regardless of whether it's being encrypted across the network. This is a common problem for verifying encryption. This research verifies encrypted IP datagrams by configuring a single Bluefield-2 to send encrypted traffic across the network. If the other Bluefield-2 is able to see the ESP headers on the IP datagrams sent from the other side the network, it is reasonable to assume that the IP traffic is properly encrypted. This verification method is assumed to be sufficient for the purposes of this research, but other verification methods should be used in switched networks.

In switched networks, a third workstation set-up with port mirroring would be required to verify data is properly encrypted. The third workstation would be capable of monitoring network traffic and verifying encrypted IP datagrams. Alternatively, a software bridge on a third Bluefield-2 could be used to capture traffic between two other Bluefield-2s.

5. Results and analysis

5.1 Factor screening

The full factorial design described in Section 4.4.1 was performed on each of the application scenarios with three replicated of each trial. In total, 190 (16 treatments x 4 scenarios x 3 replicants) iPerf3 throughput tests were performed. The residuals of the collected dataset were *nonnormal*. Therefore, ANOVA is *not* an adequate model for the data set.

5.1.1 Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric alternative to ANOVA for situations where the normality assumption is unjustified (Montgomery, 2019). Kruskal-Wallis uses an F-test analysis of variance that does not require normal residuals. This test is a good fit for the data collected for the full factorial experiment in this research.

Table 2: Full factorial design: Kruskal-Wallis test results

Treatment / Factor	Plain Text	Plain Text Offload	IPsec	IPsec Offload
Thread	90.0%	N.S.	90.0%	90.0%
MSS	99.9%	99.9%	99.9%	99.9%

Table 2. lists the results of the Kruskal-Wallis test performed on the full factorial design dataset. In all the application scenarios, the response variable is dependent on MSS according to a 99.9% confidence interval. The results of the Kruskal-Wallis test also indicate the three of the four application scenarios are dependent on the number of iPerf3 threads used in the throughput test according to a 90% confidence interval.

5.2 Hardware accelerator characterization

This section seeks to characterize the performance of the Bluefield-2's hardware accelerators. This is accomplished by running additional throughput tests with varying MSS and thread factor levels. The direction of throughput tests and the performance setting of workstation CPUs are held constant for the remaining experiments because the Kruskal-Wallis test in Section 5.1 determined that they do not significantly affect the response variable. Workstation 3 and workstation 4 act as the client and server for the following throughput tests. The workstation CPUs are configured in *performance* mode.

The performance characterization experiment described in Section 4.4.2 was performed on each of the application scenarios with three replicates of each trial. In total, 960 (10 MSS levels x 8 thread levels x 4 scenarios x 3 replicants) iPerf3 throughput tests were performed.

5.2.1 Sending plain text without hardware acceleration

The baseline performance of the Bluefield-2 peaks just above 15 Gb/s as shown by the results in Figure 3-A. Increasing the MSS during the iPerf3 throughput tests increases the average throughput of the system, whereas increasing the thread count decreases the average throughput. It is possible that the virtual bridge on the Bluefield-2 is saturated. Additional threads slow system performance due to additional context switching.

5.2.2 Sending plain text with hardware acceleration

Hardware acceleration on the Bluefield-2 increases the throughput to over 30 Gb/s as seen in Figure 3-B. The overall throughput increases by over 100% from the baseline throughput, and the maximum performance recorded during these experiments used 7 iPerf3 threads and had an MSS of 8900. The performance improvement seen in this experiment demonstrates one of the benefits of hardware offloading network traffic.

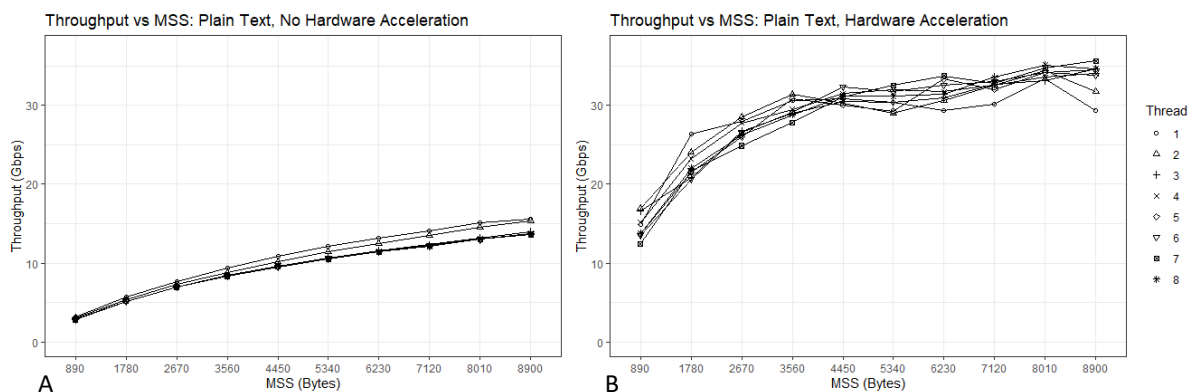


Figure 3: Sending plain text (A) No hardware acceleration (B) Hardware acceleration

5.2.3 Sending encrypted text (IPsec) without hardware acceleration

IPsec encryption significantly reduces network performance when Libreswan is run on the host CPUs. Figure 4-A. lists the performance results when IPsec is implemented without hardware acceleration. The throughput

increases as MSS and thread count increase. The average throughput peaks just above 6.5 Gb/s which is a 130% loss in performance from the baseline performance.

5.2.4 Sending encrypted text (IPsec) with hardware acceleration

Lastly, Figure 4-B. illustrates network performance when IPsec operations are offloaded to the hardware accelerators on the Bluefield-2s. When the MSS is 8900 Bytes, the average throughput of the network peaks at nearly 16 Gb/s. The hardware accelerator of the Bluefield-2 increases the average network throughput by over 140% from performance of implementing IPsec without hardware offload, and is similar to the baseline performance of the Bluefield-2.

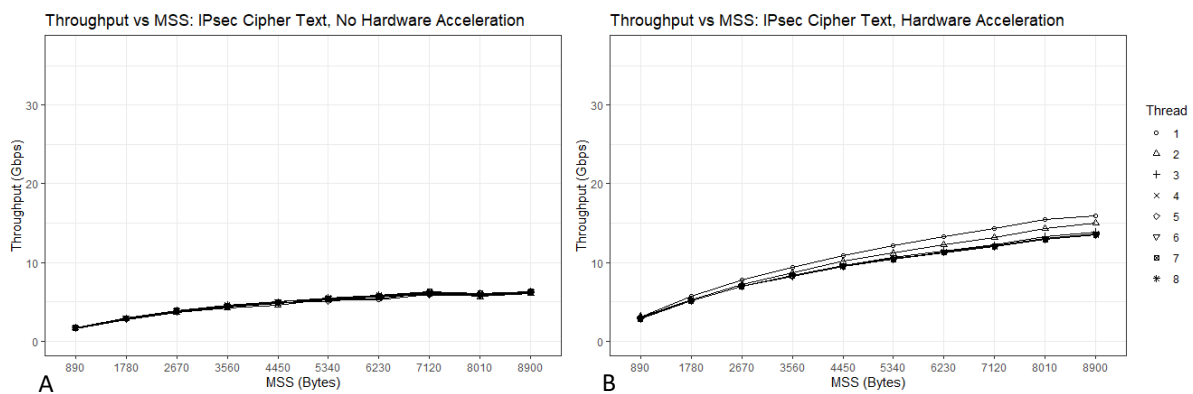


Figure 4: Sending encrypted text (IPsec) (A) No hardware acceleration (B) Hardware acceleration

5.2.5 Combined Results

Figure 5. illustrates the difference in throughput achieved by each treatment. The baseline plain text throughput is nearly identical to that of the hardware accelerated IPsec encryption. Further, hardware acceleration of plain text traffic more than doubles the throughput in the network.

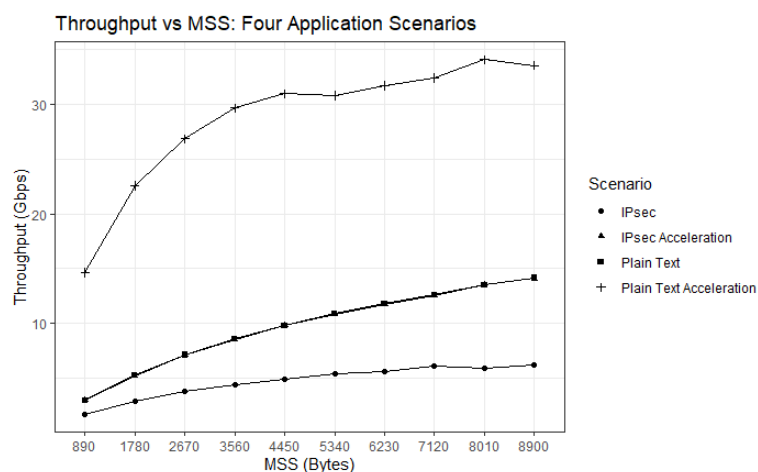


Figure 5: Comparison of average throughput

Figure 6. compares the average throughput measured during of the four application scenarios of this research and illustrates the increase in performance as a result of offloading traffic through the Bluefield-2's hardware accelerators. Hardware acceleration significantly improved the throughput for both plain text and IPsec encrypted traffic based on a 99.9% confidence interval.

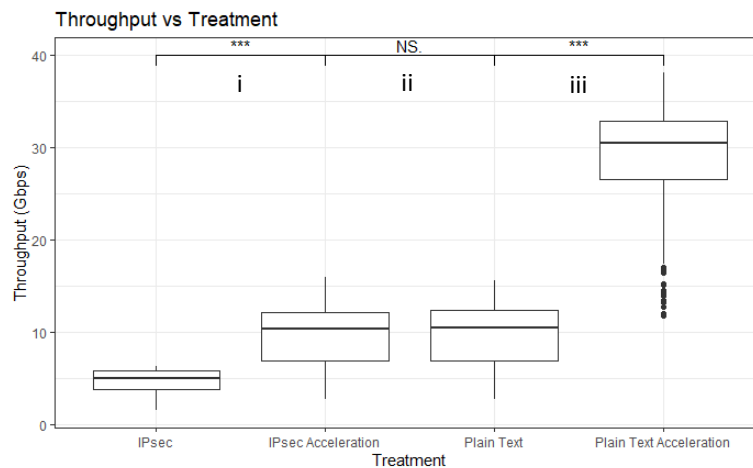


Figure 6: Combined Results (i and iii) Hardware acceleration significantly improves throughput for Plain Text and IPsec (ii) Baseline and hardware accelerated IPsec throughput are statistically the same

6. Conclusion

Previous research identified vulnerabilities in native InfiniBand's clear text key exchange and proposed changes to the IBA specification for a secure key exchange scheme using GCM encryption. The proposed key exchange scheme degraded InfiniBand network performance by up to 12%, and increased the computational load placed on CPUs. This research demonstrates the performance benefit of offloading the encryption of Ethernet traffic to hardware accelerators like those found on the Bluefield-2 DPU. Bluefield-2 DPUs are capable of swapping Ethernet and InfiniBand at the link-layer. Open-source Ethernet tools like iPerf3 are well documented. As a result, this research uses Ethernet to demonstrate the Bluefield-2s capability to hardware accelerate IPsec encryption. Initial tests using iPerf3 showed that the Bluefield-2 DPUs were capable of offloading IPsec encryption to their hardware accelerators. Subsequent throughput tests determined that the Bluefield-2 DPU IPsec/TLS specific hardware accelerators are capable of encrypting IPsec datagrams at nearly 16 Gb/s while using TCP and Ethernet at the transport and data link layers respectively. Therefore, the hardware accelerators realize a significant performance improvement when compared to the throughput of encrypting Ethernet traffic using Libreswan on the workstation's CPUs. The Bluefield-2 DPU provides an effective means for improving network performance and providing confidentiality, integrity, and authentication to Ethernet traffic. Further research is required to characterize the capability of the Bluefield-2 DPU to accelerate DMA traffic in hardware, and measure the difference in CPU utilization as a result of kernel bypass.

6.1 Future work

SmartNICs like the Nvidia-Mellanox Connect-X 6 network adapter enable the execution of extended Berkeley Packet Filters (eBPF) on the NIC itself using Express Data Path (XDP). Thus, eBPF provides a framework to enable offloading for a broad set of applications (Hohlfeld *et al.*, 2019). eBPF/XDP-based offloading from the user-space is an interesting research area. Future research should investigate the Bluefield-2s ability to offload encryption to the hardware of the Connect-X 6 NIC itself.

Additionally, DMA bypasses the OS kernel of the host and Bluefield-2s and prevents kernel space applications like TCP dump, Wireshark, and ntopng from capturing most network traffic. Therefore, monitoring InfiniBand traffic requires user space applications that use InfiniBand verbs to facilitate packet filtering. This research uses OpenFlow rules and Open vSwitch to offload encryption. These applications also do not work with InfiniBand Traffic. Therefore, future research should investigate user space applications that will enable packet filtering and encryption of RDMA InfiniBand traffic.

Acknowledgements

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, of the U.S. Government. This document has been approved for public release; distribution unlimited, case #88ABW-2021-0849.

References

- Hintze, K. (AFIT/USAF) (2021) 'Infiniband Network Monitoring: Challenges and Possibilities', *M.S. thesis, Air Force Institute of Technology*.
- Hohlfeld, O. et al. (2019) 'Demystifying the Performance of XDP BPF', *Proceedings of the 2019 IEEE Conference on Network Softwarization: Unleashing the Power of Network Softwarization, NetSoft 2019*, pp. 208–212. doi: 10.1109/NETSOFT.2019.8806651.
- InfiniBand Trade Association. Available at: <https://www.infinibandta.org/> (Accessed: 7 June 2021).
- Kurose, J. and Ross, K. (2017) *Computer Networking A Top-Down Approach*. 7th edn. Pearson.
- Lee, M. and Kim, E. J. (2007) 'A comprehensive framework for enhancing security in InfiniBand architecture', *IEEE Transactions on Parallel and Distributed Systems*, 18(10), pp. 1393–1406. doi: 10.1109/TPDS.2007.1079.
- Liu, J. et al. (2021) 'Performance Characteristics of the BlueField-2 SmartNIC'. Available at: <http://arxiv.org/abs/2105.06619>.
- Mireles, L. (AFIT/USAF) (2020) 'Implications and Limitations of Securing an InfiniBand Network', *M.S. thesis, Air Force Institute of Technology*.
- Montgomery, D. (2019) *Design and Analysis of Experiments*. 10th edn. Wiley.
- 'NVIDIA BLUEFIELD-2 DPU DATA CENTER INFRASTRUCTURE ON A CHIP' (2021). Nvidia.
- Pfister, G. F. (2001) 'An introduction to the InfiniBand™ architecture', *High Performance Mass Storage and Parallel I/O: Technologies and Applications*, pp. 617–632. doi: 10.1109/9780470544839.ch42.
- Rothenberger, B. et al. (2021) 'ReDMArk : Bypassing RDMA Security Mechanisms', *Usenix Security'20*.
- Schmitt, D. (AFIT/USAF) (2019) 'A Framework for Cyber Vulnerability Assessments of InfiniBand Networks', *M.S. thesis, Air Force Institute of Technology*.
- Strohmaier, E. et al. (2021) *TOP 500 The List*. Available at: <https://top500.org/statistics/list/> (Accessed: 7 June 2021).

Malware Binary Image Classification Using Convolutional Neural Networks

John Kiger, Shen-Shyang Ho and Vahid Heydari

Rowan University, Glassboro, NJ, USA

kigerj19@students.rowan.edu

hos@rowan.edu

heydari@rowan.edu

Abstract: The persistent shortage of cybersecurity professionals combined with enterprise networks tasked with processing more data than ever before has led many cybersecurity experts to consider automating some of the most common and time-consuming security tasks using machine learning. One of these cybersecurity tasks where machine learning may prove advantageous is malware analysis and classification. To evade traditional detection techniques, malware developers are creating more complex malware. This is achieved through more advanced methods of code obfuscation and conducting more sophisticated attacks. This can make the manual process of analyzing malware an infinitely more complex task. Furthermore, the proliferation of malicious files and new malware signatures increases year by year. As of March 2020, the total number of new malware detections worldwide amounted to 677.66 million programs. In 2020, there was a 35.4% increase in new malware variants over the previous year. This paper examines the viability of classifying malware binaries represented as fixed-size grayscale using convolutional neural networks. Several Convolutional Neural Network (CNN) architectures are evaluated on multiple performance metrics to analyze their effectiveness at solving this classification problem.

Keywords: Malware Analysis, Malware Classification, Malware Visualization, Convolutional Neural Networks, Deep Learning

1. Introduction

The shortage of professional talent in the cybersecurity workforce is an issue that continues to persist as more businesses recognize the value of cybersecurity for their organization. In their 2021 Cybersecurity Workforce Study, the cybersecurity professional organization (ISC)2 reported an estimated 2.72 million global shortage of cybersecurity professionals (ISC2, 2021). Simultaneously, malware continues to spread at an unprecedented rate, with hundreds of thousands of new signatures detected every day. In 2020, Kaspersky's detection systems discovered an average of 360,000 new malicious files every day over the past 12 months—18,000 more than the previous year (a 5.2% increase) and up from 346,000 in 2018. 60.2% of those malicious files were non-specific Trojans. In general, the percentage of Trojans detected increased by 40.5% when compared to the previous year (Kaspersky Lab, 2021). Observing these trends gives the primary motivation for seeking automated solutions for malware analysis, particularly in the areas where analysis is done manually. Traditionally, malware analysis is conducted using one, or a combination of, static and dynamic malware analysis methods. However, the emergence of fileless malware has prompted a new method of memory-based analysis. Here we will provide a high-level description of current malware analysis methods to provide a deeper understanding of the current landscape.

2. Static Malware Analysis

Static malware analysis is a signature-based approach that entails deriving information about a potentially malicious executable file without the executing file. This is done through a variety of methods that attempt to derive and enumerate any signatures that classify a given executable file as malicious. Traditional static analysis methods are basic and do not offer deeper insights into a malicious file's behavior. Furthermore, sophisticated malware easily eludes traditional static malware analysis through increasingly complex code obfuscation techniques. However, these methods are the least computationally intensive and time-consuming to conduct. Therefore, static malware analysis is often the first method utilized when investigating an executable file. Typically, static malware analysis is achieved through these methods:

- Checking the executable file's hash against a known malware database, such as Virustotal. These databases are open-source and are usually evaded by serious malware developers.
- Using the 'strings' command on the executable file in a command-line interface to find any malicious keywords contained in the file's strings. This method is not effective on encrypted versions of malware.

- Manually decompiling the file in a disassembler, such as IDA, to analyze an executable file's control flow. This method requires subject matter expertise and may become intractably complex if the malicious code is severely obfuscated.

3. Dynamic Malware Analysis

Dynamic malware analysis is a behavior-based approach that entails executing a potentially malicious file in a virtualized sandbox environment and attempting to gain insights into a file's runtime behaviour through closely monitored observation. This method is the most computationally intensive of all the analysis methods due to the need to run the executable file in a virtualized sandbox. This prevents any damage a malicious file may cause to a system. Furthermore, this method may prove to be the most time-consuming as many malware files significantly delay the execution of their malicious code to circumvent antivirus software. Finally, sophisticated malware often utilizes techniques to determine if it is running in a virtualized environment and if so, will appear benign to this method of analysis. While this method is subject to be circumvented, if successful, significant insights on a malware's behaviour can be derived from the following methods:

- Monitoring malicious network traffic activities through tools, such as Wireshark. This allows the capture and deep analysis of all traffic originating from the executable file.
- Process monitoring can also provide insights into how a potentially malicious file interacts with the host system. The virtualized environment can capture calls to the host device's file system and registry to monitor any malicious activity, such as deleting or manipulating any important system resources.

4. Memory Malware Analysis

Most malware persists through an executable file or latching itself to a critical system resource. The most sophisticated and devastating of these are Advanced Persistent Threats (APT). However, stealthier attack vectors that attempt to elude traditional dynamic analysis methods, are classified as Advanced Volatile Threats (AVT). Fileless malware is classified as an AVT since it is malicious software that exclusively operates as an artifact in a system's volatile memory or RAM. Like dynamic analysis, this method requires a virtual environment with allocated RAM to capture and dump a memory image and analyze artifacts that include:

- Process lists and associated threads
- Networking information and interfaces (TCP/UDP)
- Kernel modules, including hidden ones
- Bash and command history
- System calls
- Kernel hooks

5. Related Research

This section includes a brief review of some previous image-based automated malware classification approaches.

5.1 Deep learning at the shallow end: Malware classification for non-domain experts

Le et al. (2018) presented a deep learning-based malware classification approach that required no expert domain knowledge and is based on a pure data drive approach for complex pattern and feature identification. Acknowledges the lack of domain expertise when handling digital evidence in law enforcement. The research specifically cites how Artificial Intelligence (AI) can aid digital investigators. AI can expedite the investigative process and ultimately reduce the case backlog while avoiding bias and prejudice (James and Gladyshev, 2013). The contribution of this work includes a deep learning model which achieves a 98.2% accuracy in classifying raw binary files into one of 9 classes of malware. The one-dimensional representation of a raw binary file is like the image representation of a raw binary file in the work of Nataraj et al. (2011). The sequential representation allowed Le et al. (2018) to apply a Convolutional Neural Network – Bi Long Short Term Memory architecture (CNN-BiLSTM) which helped achieve better performance than using a CNN model alone.

5.2 Malware Detection by Eating a Whole EXE

The most significant finding of this work from Raff et al. (2017) is how the batch normalization method hindered the learning process for all their proposed models, in all their approaches, on all the software platforms they experimented on. When given the context of analyzing binary executables, their hypothesis is well-reasoned.

We hypothesize that batch norm's ineffectiveness in our model is a product of training on binary executables. The majority of contemporary deep learning research, including batch-normalization, has

been done in the image and signal processing domains, with natural language a close second. In all of these domains, the nature of data is relatively consistent. In contrast, our binary data present a novel multi-modal nature of the byte values that can have drastically different meanings depending on the location, ranging from ASCII text, code, structured data, or even images stored for the icon. Our hypothesis is that this multi-modal nature produces multiple modes of activation, which violates the primary assumptions of batch-normalization, causing degraded performance. (Raff et al., 2017).

6. Objective

This work aims to research and perform practical experiments to determine the effectiveness of using a deep learning approach, such as a convolutional neural network, for an automated static malware detection and classification scheme. To investigate the effectiveness of different convolutional neural network (CNN) architectures for this purpose, multiple CNNs were tested, and their performance at solving this multiclass image classification problem was compared. Training time was used as a rudimentary measure of computational complexity to determine each network's applicability to run on a device with strict resource limitations, such as an IoT device. The ability to perform these tasks on a low-powered edge device was taken under consideration, as the proliferation of IoT-based malware has been seeing a significant increase. In 2019, SonicWall Capture Labs threat researchers recorded 34.3 million IoT malware attacks. In 2020, that number rose to 56.9 million, a 66% increase (SonicWall, 2021). Traditional performance metrics used for neural networks, such as accuracy, loss, F1 scores, precision, and recall, were used to determine the strength of each network architecture.

Table 1: Maling classes, and their respective sample sizes

#	Class Name	Malware Type	Samples
1	Adailer.C	Dialer	122
2	Agent.FYI	Backdoor	116
3	Allaple.A	Worm	2949
4	Allaple.L	Worm	1591
5	Alueron.gen!J	Worm	198
6	Autorun.K	Worm:AutoIT	106
7	C2LOP.P	Trojan	146
8	C2LOP.gen!g	Trojan	200
9	Dialplatform.B	Dialer	177
10	Dontovo.A	Trojan Downloader	162
11	Fakerean	Rogue	381
12	Instantaccess	Dialer	431
13	Lolyda.AA1	Password Stealer	213
14	Lolyda.AA2	Password Stealer	184
15	Lolyda.AA3	Password Stealer	123
16	Lolyda.AT	Password Stealer	159
17	Malex.gen!J	Trojan	136
18	Obfuscator.AD	Trojan Downloader	142
19	Rbot!gen	Backdoor	158
20	Skintrim.N	Trojan	80
21	Swizzor.gen!E	Trojan Downloader	128
22	Swizzor.gen!I	Trojan Downloader	132
23	VB.AT	Worm	408
24	Wintrim.BX	Trojan Downloader	97
25	Yuner.A	Worm	800

6.1 Dataset Description

The Maling dataset (Nataraj et al., 2011) is used in our investigation. This dataset contains 9,339 unique malware binary files, also known as portable executable (PE) files, which have been converted to PNG images. Each image in the dataset belongs to one of 25 total classes of malware. Each CNN was tasked with identifying the unique features of each malware class during training and accurately predicting the class name of a given malware binary image during validation. Table 1 shows the 25 classes and their respective number of instances.

6.2 Maling Class Distribution and Normalization

We observe that nearly half of the dataset is comprised of only two malware classes, Allaple.A and Allaple.L. This can prove to be a severe issue when training the neural networks, so all of the class weights of the training data were normalized using the `class_weight.compute_class_weight` function from SciKit Learn's `utils` library. While this distribution may not be optimal for a research environment, it is a more realistically distributed dataset. In

the context of malware distribution, certain classes of malware can proliferate at an exponentially higher rate than other malware.

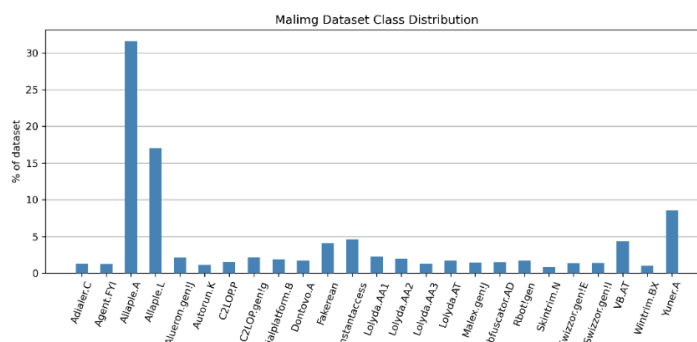


Figure 1: A bar chart demonstrating the distribution of the various classes in the Maling dataset

6.3 Observation of a Malware Image Sample

To understand the structure of a Windows Portable Executable (PE) file, and to understand how these executable binaries were converted to images, we inspect a single, unedited sample from the Maling dataset. A PE file consists of the following sections, which can be seen visually in Figure 1:

1. `.text` – The default code section which contains the executable code.
2. `.rdata` – The default read-only data section. String literal and C++/COM vtables are an example of items found in this section.
3. `.data` – The default read/write data sections. Global variables typically are stored here.
4. `.rsrc` – The section that contains all the resources of the module. This also includes icons that an application may use. (Microsoft, n.d.)

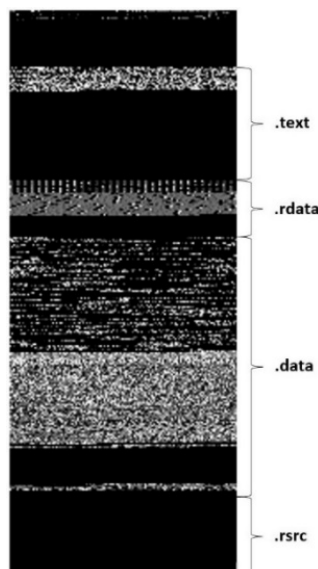


Figure 2: The code sections of a Dontovo.A Trojan

A given malware binary is read as a vector of 8-bit unsigned integers and then organized into a 2D array. This can be visualized as a grayscale image in the range [0,255] (0: black, 255: white). The width of the image is fixed, and the height is allowed to vary depending on the file size (Nataraj et al., 2011). In Figure 2, we notice multiple solid black sections, specifically in the `.text` section. This may represent uninitialized code; however, it is more likely to represent a common code obfuscation technique known as zero padding. The benefit of using a learning-based static analysis method is that code obfuscation techniques like this are generally ignored by each CNN.

6.4 Dataset Sample After Pre-Processing

The following visual data sample was compiled after pre-processing all images in the data set into 64x64 grayscale images. Not only does this compression retain the unique visual features that distinguish each malware class, but the image space occupied by zero padding is highly compressed and, therefore, less relevant as a significant feature of the data. Visual similarities between random samples of the same class, as well as visual similarities between similar classes are observed in Figure 3.

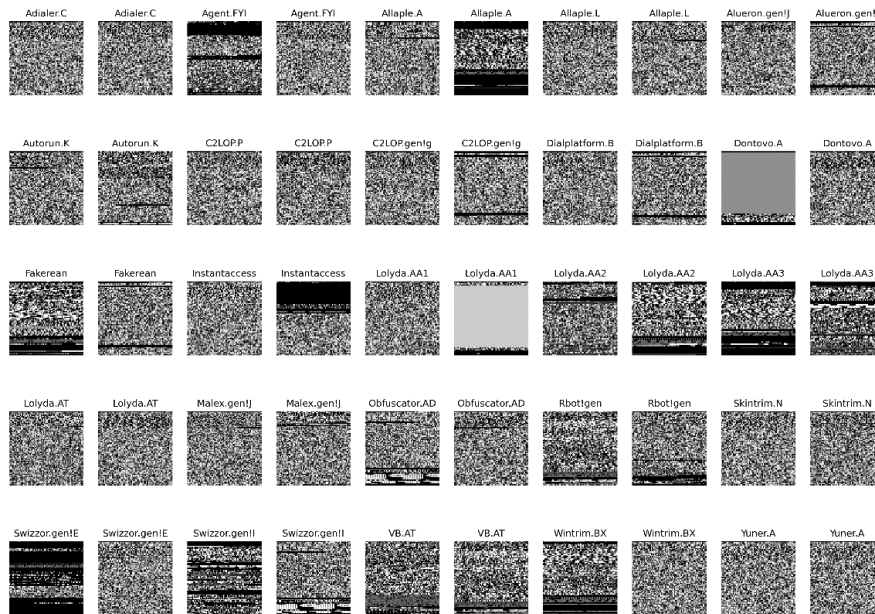


Figure 3: Samples from each class of malware from the Maling dataset after pre-processing

7. Software and Hardware Used

7.1 Software

All experiments were done in a Python3 Jupyter notebook. The machine learning libraries used were TensorFlow, Keras, and SciKit Learn. All visualizations were made using Matplotlib, and Seaborn. Other functionality, such as data processing and file I/O was accomplished using the NumPy, OS, Panda, and Collections Python libraries.

7.2 Hardware

All experiments were conducted on a 5th generation Lenovo ThinkPad X1 Carbon with an Intel Core i7-7600U with four logical cores with a 2.80Ghz clock speed and 16GB of RAM. There was no GPU acceleration utilized in these experiments.

8. CNN Performance Comparison

The performance metrics used to compare the five CNNs tested during experimentation were:

1. Total Training Time
2. Average Training time per Epoch
3. Total Accuracy
4. Total Loss
5. Weighted Average F1 Score
6. Weighted Average Precision
7. Weighted Average Recall

The purpose of our experiments was to compare the accuracy and overall strength of each network and test each network's applicability for running natively on a lower-powered device, such as an IoT or network device. One of static malware analysis strengths is the low computational resources required to analyze a potentially malicious file. Therefore, the goal of these experiments was to create a CNN that would not only classify malware accurately but with relatively great efficiency and low computational overhead. The time-based performance metrics serve as a rudimentary measure of each network's efficiency. All models used a 70/30 training/validation

split during testing. Furthermore, all models used categorical cross-entropy as the loss function and the Adam optimizer on their compilation layers. Therefore, these constants are omitted when detailing each model's architecture.

8.1 Comparison of Convolutional Model Architectures

A total of five CNN models were evaluated during our experiments. These models are split into two distinct subcategories. Models 1 and 2 use batch normalization as the normalization function, and Models 3, 4 and 5 use Local Response Normalization as the normalization function. Models 1 and 2 retain the same input size of 64x64x1, with model 2 having double the number of convolutional layers in an experiment to decrease training time between epochs. Models 3, 4 and 5 were designed to easily scale to different input sizes and designed to use a new normalization method. The latter design change was done to test the hypothesis of Raff et al. (2017) who cited the failure of batch normalization in their experiments, while the former was done to test the performance impact of using different input sizes.

8.1.1 A Comment on Normalization Functions

As Raff et al. (2017) cited in their work, batch normalization failed in their experiments by severely hindering the learning process on malware binaries. In their hypothesis they cited that byte values can have drastically different meanings depending on the location, ranging from ASCII text, code, structured data, or even images stored for the icon (Raff et al., 2017). This reasoning is consistent when given the context of how the batch normalization function operates and given that the malware image binaries contain structured contextual data, which was detailed in Section 4.3. In an article titled *Difference between Local Response Normalization and Batch Normalization*, Anwar, A. (2021) details the process of the batch normalization function. Before being fed to the activation function, the output of the hidden neurons is processed as by completing the following steps:

1. The entire batch B is normalized to a zero mean and unit variance
2. The mean is calculated for the entire mini-batch output.
3. The variance is calculated for the entire mini-batch output.
4. The mini-batch output is normalized by subtracting the mean and dividing by the variance.
5. Two trainable parameters are introduced, Gamma, the scale variable and Beta, the shift variable which scales and shifts the normalized mini-batch output.
6. The scaled and shifted normalized mini-batch output is fed to the activation function. (Anwar, 2021)

We hypothesize that the scaling and shifting of the batch data done during the normalization function is the reason why batch normalization may prove to hinder the learning process for contextually sensitive data, such as an image representation of a binary file. When the image is scaled and shifted during the batch normalization function, the feature's context is lost. For example, say a given sample contains a series of bits that represent a malicious API call in the .text section of a Windows PE file, if the image is scaled and shifted to where this malicious string is moved to a different section of the PE file structure, it's meaning, and context are lost to the learning algorithm.

Local response normalization does not use scaling and shifting but rather uses the concept of lateral inhibition. Lateral inhibition is a concept from neurobiology where stimulated neurons inhibit the activity of neighbouring neurons. For local response normalization, this concept of lateral inhibition is used to conduct local contrast enhancement, so the local maximum pixel values are used as the excitation for the next layers (Anwar, 2021). Local response normalization, which relies on the local context of a given pixel, rather than the shifting and scaling the image relative to the gamma and beta values in batch normalization. Local response normalization function was considered as the most appropriate normalization function over batch normalization when considering the contextual nature of the binary data. After experimenting with model 1 and 2, which used batch normalization, the remaining models utilized local response normalization. Once its effectiveness was verified in our experimentation, we then moved to experiment with different input sizes for our training and testing data. This is also the reason why models 3, 4 and 5 were designed to scale well with different sized input data.

Table 2: Comparison of the model architectures tested

Architectural Comparison					
Model No.	Input Dimension	Convolutional Layers	Pooling	Normalization Function	Dense Layers
1	64x64x1	30 Filters, 3x3 Kernel Size, Relu Activation 15 Filters, 3x3 Kernel Size, Relu Activation	Max Pooling 2x2 Pool Size	Batch Normalization	128 Units – Relu 50 Units – Relu 25 Units – Softmax
2	64x64x1	32 Filters, 3x3 Kernel Size, Relu Activation 32 Filters, 3x3 Kernel Size, Relu Activation 64 Filters, 3x3 Kernel Size, Relu Activation 128 Filters, 3x3 Kernel Size, Relu Activation	Max Pooling 2x2 Pool Size	Batch Normalization	256 Units – Relu 128 Units – Relu 25 Units – Softmax
3	128x128x1	50 Filters, 5x5 Kernel Size, Relu Activation 70 Filters, 3x3 Kernel Size, Relu Activation 70 Filters, 3x3 Kernel Size, Relu Activation	Max Pooling 2x2 Pool Size 1 Stride	Local Response Normalization	256 Units – Relu 25 Units – Softmax
4	64x64x1	25 Filters, 5x5 Kernel Size, Relu Activation 35 Filters, 3x3 Kernel Size, Relu Activation 35 Filters, 3x3 Kernel Size, Relu Activation	Max Pooling 2x2 Pool Size 1 Stride	Local Response Normalization	128 Units – Relu 25 Units – Softmax
5	32x32x1	15 Filters, 5x5 Kernel Size, Relu Activation 25 Filters, 3x3 Kernel Size, Relu Activation 25 Filters, 3x3 Kernel Size, Relu Activation	Max Pooling 2x2 Pool Size 1 Stride	Local Response Normalization	64 Units – Relu 25 Units – Softmax

8.2 Performance Results

Table 3: Performance results of models tested

Performance Comparison (<i>best results are bolded</i>)								
Model No.	Total Epochs	Total Training Time	Average Training Time per Epoch	Total Accuracy	Total Loss	Average Weighted F1 Score	Average Weighted Precision	Average Weighted Recall
1	25	9.23 minutes	22.16 seconds	85.22%	31.41%	0.84	0.85	0.85
2	25	5.87 minutes	14.08 seconds	88.69%	16.17%	0.88	0.88	0.89
3	10	3.59 hours	21.52 minutes	95.07%	29.98%	0.94	0.94	0.95
4	10	17.38 minutes	1.74 minutes	93.97%	23.12%	0.93	0.93	0.94
5	50	11.82 minutes	14.18 seconds	77.59%	135.82%	0.77	0.77	0.78

8.2.1 Model 1 and 2 Performance Comparison

Models 1 and 2 use very similar to architectures, both use a 64x64x1 input size, both trained for 25 epochs, and both use batch normalization as their normalization function. Due to their similarity, their performance results will be compared directly. Outside of model 5, which used a significantly reduced image input size, model 1 was the worst performing model regarding accuracy. While training only took under 10 minutes, the accuracy gave

room for improvement. Model 2 was designed to expand on model 1. First, it was designed to increase accuracy and reduce training time. While model 2 achieved both goals, with a 3.47% increase in accuracy with a 3.36-minute reduction in training time, there was concerns about its strength as a model.

In Figure 4, there was a significant deviation between model 2's validation and training accuracy starting at the 20th epoch. This is not ideal as the goal is to have the validation curve follow the training curve as closely as possible. Upon further investigation, we theorized that this could have been caused by using batch normalization or the architecture of models 1 and 2.

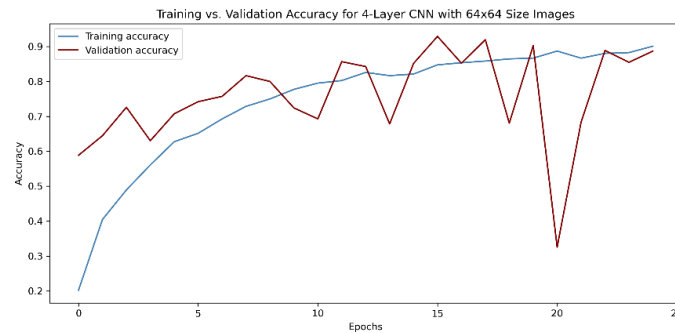


Figure 4: Training vs. validation accuracy for model 2

8.2.2 Model 3, 4 and 5 Performance Comparison

Models 1 and 2 exhibited significant deviation between training and validation on accuracy and loss curves. This behaviour prompted a significant restructuring of the subsequent networks, prioritizing scalability to easily test different input sizes and the switch from batch normalization to local response normalization. The first model to test this new architecture was model 3, which used a 128x128x1 input size. Doubling the input shape over the previous two model's 64x64x1 input shape increased the accuracy, with the best-recorded accuracy for model 3 being 95.07%. However, the training time was significantly longer, at nearly four hours.

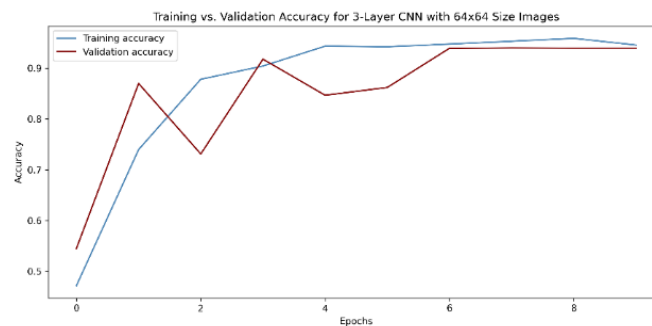


Figure 5: Training vs. validation accuracy for model 4

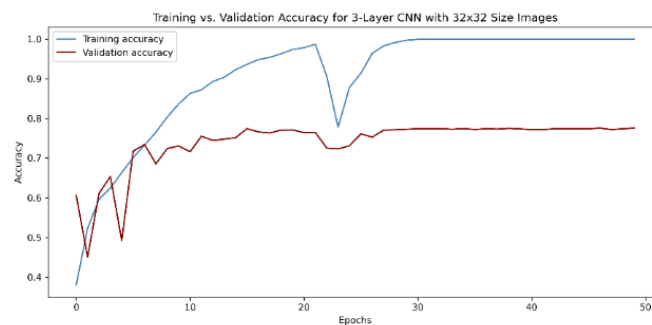


Figure 6: Training vs. validation accuracy for model 5

The lengthy training time of model 3 prompted us to revert to the previous 64x64x1 input shape in model 4 and scale the network down accordingly. Model 4 trained for the same number of epochs as model 3 and its performance results were encouraging. Model 4 reduced the training time of model 3 by 92%, while only

decreasing accuracy by 1.2%. This was a preferred trade-off, and a significant reduction in training time was desired. Curious about the effects of reducing the input size further, we scaled down the images to a 32x32x1 input size in model 5, with the same downscaling of the model as conducted in model 4. Observing the results for model 5 in Table 3, the effect of scaling the images down to such a reduced image size had a negative effect on model 5's performance. This is also displayed in training vs. validation accuracy plots of both models. Figure 5 plots model 4's training vs. validation accuracy curve. In Figure 5, the validation curve closely follows the training curve through all of model 4's 10 epochs. In Figure 6, model 5's training vs. validation accuracy curve shows a significant deviation from the training curve, visually displaying the issues with model 5.

Furthermore, model 5 trained for a total of 50 epochs to boost its accuracy, which was only 77.59%. This eliminated any training time reduction desired by reducing the input size, with only a modest reduction of 5.56 minutes of training time over model 4 but a 16.38% reduction in accuracy. This significant performance loss can be attributed to the severe reduction in image size. This reduced image size lost many unique class features, contributing to the inter-class confusion in similar classes. In model 5's confusion matrix, the similar classes Allapple.A and Allapple.L appeared to the network as the same class. Any feature that could uniquely identify these classes was lost when reducing the image size to 32x32x1. This confusion is displayed visually in Figure 7 which plots model 5's confusion matrix.

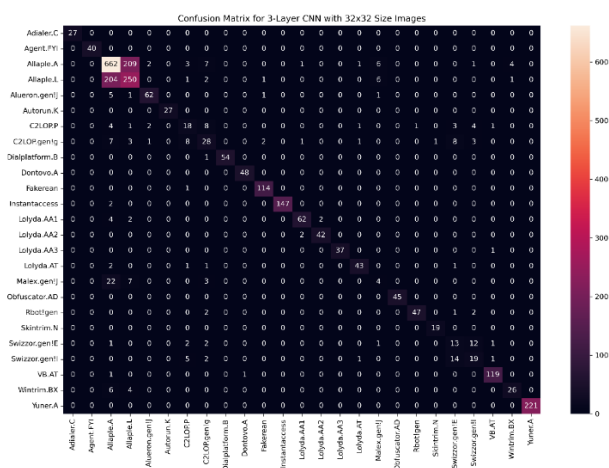


Figure 7: Model 5 confusion matrix

9. Conclusion

Some issues regarding the future viability of this method were not covered in this work. An infinitely complex feature set is the most prescient issue concerning automating malware classification. Since unique malware signatures continue to grow, this will add more classes to an already large dataset. This complexity could prevent learning and cause severe inter-class confusion when presented to a learning model. This confusion could lead to misclassification, which can severely hinder the accuracy of a trained model. The dataset we have utilized is only an extremely small subset of the millions of relevant malware classes currently distributed online. To realistically classify malware of all relevant malware signatures, rather than a small subset, future research must be conducted into generalizing the classes to a level conducive to effective and fast learning.

Another relevant issue that must be addressed is the difficulty of gathering a large malware dataset, which is even more challenging when gathering a large dataset with malicious and benign files. This data aggregation issue does not only affect our research, but it affects all malware researchers. For example, our dataset was not as easily accessible as one might expect, and it did not contain the original malware binaries. This may be due to file hosting services, such as Dropbox, automatically flagging files containing malware. We would have liked to obtain a similar dataset that contained the original PE files, where we would have conducted our own conversion to grayscale images. The PNG images in the Malimg data set were not true grayscale images, having a color depth of 3, rather than 1. This had to be addressed during preprocessing of our dataset, where all the images were converted to true grayscale using Python's image library. Until malware analysts and researchers have an open and easily-accessible dataset to conduct experiments with, this area of research will continue to fall behind other fields of machine learning.

The final issue which must be considered is performance. Training a convolutional neural network is a computationally-intensive process. For example, our third model required nearly four hours of training on a mid-level laptop with no GPU acceleration. Ideally, we would like to run a production version of this automated system on a network edge device. This would be the most conducive application of our proposed malware detection system since static malware analysis techniques are the first line of defense against malicious threats. Edge devices typically have limited computational resources. Providing a neural network that can run natively and in real-time on these edge devices is essential to achieving this goal.

Finally, what has been presented should not be considered a method to supersede or replace any of the current static malware analysis methods. However, the proposed method should be considered a novel tool in a malware analyst's skill chain. Unlike natural language processing, machine learning is not a silver bullet that solves all cybersecurity-related tasks. Instead, machine learning ought to automate a security professional's most time-consuming and repetitive tasks. Automating some of the initial stages of static malware analysis would allow malware analysts more time to gain deeper insights into the malicious file at hand. We are confident that our research shows that our proposed automated static malware analysis method can become a viable real-time malware detection method.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1753900.

References

- Anwar, A. (2021). *Difference between Local Response Normalization and Batch Normalization*. [online] Medium. Available at: <http://towardsdatascience.com/difference-between-local-response-normalization-and-batch-normalization-272308c034ac> [Accessed 3 Feb. 2022].
- Bhodia, N., Prajapati, P., Troia, F. and Stamp, M. (2019). *Transfer Learning for Image-Based Malware Classification*. Gibert, D., Mateu, C., Planes, J. and Vicens, R. (2018). Using convolutional neural networks for classification of malware represented as images. *Journal of Computer Virology and Hacking Techniques*, 15(1), pp.15–28.
- ISC2 (2021). *2021 Cybersecurity Workforce Study*. [online] www.isc2.org, ISC2, pp.24–25. Available at: <https://www.isc2.org/-/media/ISC2/Research/2021/ISC2-Cybersecurity-Workforce-Study-2021.ashx> [Accessed 4 Feb. 2022].
- Kaspersky Lab (2021). *The number of new malicious files detected every day increases by 5.2% to 360,000 in 2020*. [online] www.kaspersky.com. Available at: https://www.kaspersky.com/about/press-releases/2020_the-number-of-new-malicious-files-detected-every-day-increases-by-52-to-360000-in-2020 [Accessed 4 Feb. 2021].
- Le, Q., Boydell, O., Mac Namee, B. and Scanlon, M. (2018). Deep learning at the shallow end: Malware classification for non-domain experts. *Digital Investigation*, [online] 26, pp.S118–S126. Available at: <https://www.sciencedirect.com/science/article/pii/S1742287618302032> [Accessed 30 Jan. 2020].
- Mallet, H. (2020). *Malware Classification using Convolutional Neural Networks — Step by Step Tutorial*. [online] Medium. Available at: <https://towardsdatascience.com/malware-classification-using-convolutional-neural-networks-step-by-step-tutorial-a3e8d97122f>.
- Microsoft (n.d.). *An In-Depth Look into the Win32 Portable Executable File Format, Part 2: Figures*. [online] bytepointer.com. Available at: https://bytepointer.com/resources/pietrek_in_depth_look_into_pe_format_pt2_figures.htm [Accessed 4 Feb. 2022].
- Nataraj, L., Karthikeyan, S., Jacob, G. and Manjunath, B.S. (2011). Malware images. *Proceedings of the 8th International Symposium on Visualization for Cyber Security - VizSec '11*.
- Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B. and Nicholas, C. (2017). *Malware Detection by Eating a Whole EXE*.
- SonicWall (2021). *2021 SonicWall Cyber Threat Report*. [online] <https://www.sonicwall.com/>, Milpitas, CA: SonicWall Inc., p.58. Available at: <https://www.sonicwall.com/resources/white-papers/2021-sonicwall-cyber-threat-report/> [Accessed 4 Feb. 2022].
- Theta432 (2020a). *Theta432*. [online] www.theta432.com. Available at: <http://www.theta432.com/post/malware-analysis-part-1-static-analysis> [Accessed 3 Feb. 2022].
- Theta432 (2020b). *Theta432*. [online] www.theta432.com. Available at: <http://www.theta432.com/post/malware-analysis-series-part-2-dynamic-analysis> [Accessed 3 Feb. 2022].
- Theta432 (2020c). *Theta432*. [online] www.theta432.com. Available at: <http://www.theta432.com/post/malware-analysis-series-part3-memory-malware-analysis>.
- Véstias, M.P. (2019). A Survey of Convolutional Neural Networks on Edge with Reconfigurable Computing. *Algorithms*, 12(8), p.154.

Defending Small Satellites from Malicious Cybersecurity Threats

Banks Lin, Wayne Henry and Richard Dill

Air Force Institute of Technology, Wright-Patterson Air Force Base, USA

banks.lin@afit.edu

wayne.henry@afit.edu

richard.dill@afit.edu

Abstract: The connection between space and cyberspace domains is increasingly intertwined. Advancements in space technology, decreasing costs for satellite development, and the use of commercial off-the-shelf products present many cybersecurity challenges to space infrastructure. Additionally, space-based global critical infrastructure makes the space domain a prime target for malicious cyber threats. Software-defined radios introduce a potential attack vector for adversaries planning malicious satellite activity. This paper demonstrates how an adversary would send malicious commands via a software-defined radio to affect the integrity of the sensors on the satellite running NASA's core Flight System software. The experiment demonstrates one possible threat vector using a commercially available USRP N210 software-defined radio. The results show that well-constructed messages can be created to manipulate sensors on a target small satellite system. Identifying cybersecurity vulnerabilities like these in space systems can improve security and prevent disruptions for the global space enterprise.

Keywords: small satellite, space cybersecurity, software-defined radio, insider threat, integrity

1. Introduction

Space technology continues to march towards new exploration and innovation. The race to launch satellites to orbit is on for many nations in academia and the commercial sector. The cost and timeline to develop small satellites are appealing and are becoming increasingly common. This is especially true with developers using open-source software like NASA's core Flight System (cFS) (Wilmot and Kane, 2021). The motivation to gravitate towards small satellites is compelling; however, it introduces cybersecurity challenges for space assets and infrastructure (Tucker, 2019). Cybersecurity testing is one way to ensure the satellite in development is secure and ready for launch. From a test perspective, it is safe to assume that adversaries possess the intent, knowledge, time, and money to exploit a vulnerability.

There is an absence of cybersecurity standards for space systems. Without an international or industry standard, organizations can launch satellites that may lack protection against cyber-attacks which may cause security concerns for years to come. The "security by obscurity" model alone is obsolete (Jameel, 2020). The model depends on the complexity of the system design and implementation primary method for security (Waterman, 2019). Instead, the obscurity of satellite systems must couple with other layers of security. A cybersecurity standard for design, development, and execution for space systems enables all satellites to meet a minimum viable product capable of cyber resiliency. Maintaining a set of principles addresses the strategic and technical approach for space system stakeholders to combat space cybersecurity (Falco, 2019). The standard should also apply during the development phase, where the lifecycle of the space system integrates cybersecurity testing. Satellite development teams that add cybersecurity defenses after launch may risk unintended side-effects that may cost more in the long run. It is imperative to conduct cybersecurity assessments of satellites during development and before launch to ensure the spacecraft can be trusted and secured for the duration of the satellite's life in orbit.

Cybersecurity assessments on the ground segment are similar to ones conducted on traditional information technology network. However, there is less research on testing the security posture of the spacecraft itself. The consequence of this is failing to learn the unknown to the cyber threats that face spacecraft. Threats to the spacecraft include but are not limited to electromagnetic interference or signal jamming. This paper takes the approach to determine if it is possible to send malicious commands to small satellites with software-defined radios (SDRs). An SDR is a radio where the physical layer functions traditionally implemented in hardware are software-defined. Software configuration to the mixer, filter, amplifier, or modulator components reduces development costs. SDRs have human configurable software code to instruct the device how to modulate or demodulate the data. Anytime humans are in the loop, there is always a method to exploit the process. Demonstrating the existence of compromising the integrity of a satellite through insider threat is only the first step of a broader research objective to discover the capabilities to compromise and spoof satellites.

Section 2 details previous research on space system cybersecurity. Section 3 covers the research methodology of this experiment and presents the system under test. Section 4 discusses the procedures done to alter commands with malicious intent to the spacecraft and reveals the experiment results. Lastly, Section 6 concludes the experiment and discusses future work.

2. Related Work

The technological advancement in space is no longer limited to nation-states and large corporations. The commercial and civilian sectors are diving into the space enterprise because of the decreasing costs of technology when it was only attainable to nation-states in the past. Commercial off-the-shelf (COTS) or open-source software and hardware, inexpensive small satellites, and decreasing cost of space launches allow more organizations to be part of the space enterprise (Nussbaum and Berg, 2020). Companies and academic institutions take on more risks to push boundaries in space technologies and applications. Although the growth in the space industry is good for research and academia, it introduces side effects that require management. Open-source products can lower the cost and development time; however, system engineers must be aware of the cybersecurity challenges. NASA's cFS is the most common open-source flight software used in small satellites. Its software code is publicly available on GitHub for any entity to download. This research will assess NASA's cFS and leverage its software code as a potential attack vector. Space systems are increasingly becoming an attractive target for adversaries because of the significant impacts on their capabilities (Manulis et al., 2020). Adversaries avoid companies that have heavy experience and investment in cybersecurity. Instead, they target sectors prone to cybersecurity vulnerabilities, such as the space sector. Nation-states have launched orbital assets for anti-satellite capabilities; therefore, the likelihood of cyber-attacks targeting space systems is high (Vessels et al., 2019).

Insider threat is one of the many prevalent threats that space missions face. Insider threats can cause the most impact because these individuals possess the necessary permissions to conduct their malicious operations. A NASA report concluded that insider threats are often dishonest systems personnel, disgruntled staff members, or trusted business partners (Book, 2015).

Reputable companies and universities have conducted extensive research on the cyber resiliency of flight software on satellites. The Aerospace-Carnegie Mellon team describes cyber resilience as needing to be an integrated capability of every system, not a "bolt-on" capability or an afterthought (Wheeler et al., 2017). Their paper emphasized two approaches to conducting a security assessment for flight software: dynamic application security testing (DAST) and static application security testing (SAST). The DAST and SAST are part of a seven-step lifecycle of analysis, design, coding, testing, integration, deployment, and maintenance, shown in Figure 1.

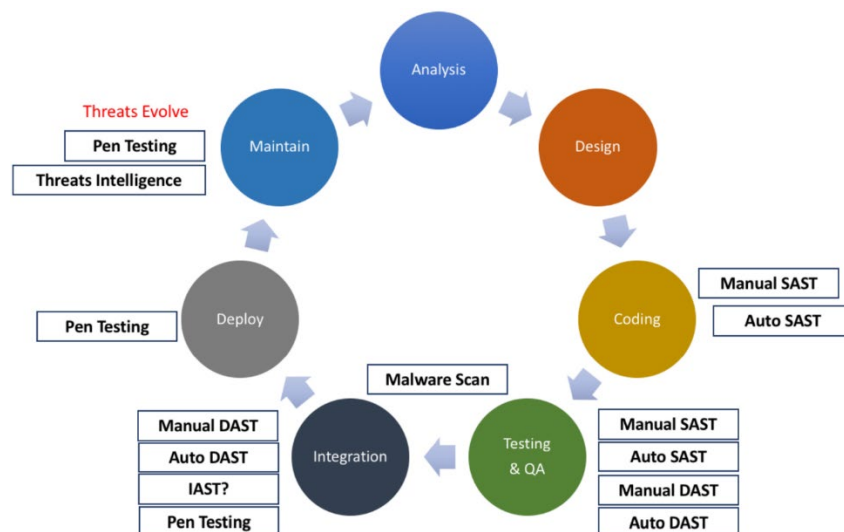


Figure 1: Software Security Assessment Lifecycle (Wheeler et al., 2017)

Hacking with SDRs is not a new concept. With low-cost SDRs and open-source software, many enthusiasts can get their hands on these tools to experiment with the internet of things (IoT). Garage door openers are common targets where attackers can capture the radio frequency (RF) signal and replay the signal from an SDR (Gupta,

2017, pp.223-263). With the right tools and skillset, enthusiasts can access hacking IoT devices that use RF communication. Hung and Vinh studied vulnerabilities in IoT devices from unknown RF protocols by analyzing the frequency and decoding the RF signal (Hung and Vinh, 2019). After decoding the recorded signal data, they rebroadcasted it to the IoT devices, performing a replay attack. This paper applies Hung and Vinh's method to satellite communication.

Hacking satellites with SDRs is also not a new concept. Lukin and Haselberger conducted a replay attack using a HackRF SDR on a Reaktor Hello World satellite (Lukin and Haselberger, 2020). In their experiment, the communication between satellite and ground station lacked proper authentication and failed to enforce encryption of the communication channel. They used HackRF as their SDR to transmit the command and open-source software called GQRX, or GNU Radio, Qt, Rx, to capture the received frame. They translated the raw data in the hexadecimal bitstream for the replay attack and sent it over Gaussian frequency-shift keying (GFSK) modulation. Their goal was to demonstrate sending commands to a satellite, simulating real communication through low-cost COTS products and open-source software.

Although Lukin and Haselberger could conduct a replay attack, they could not send commands with their own values to the satellite. This research builds on their work to send altered malicious commands to the satellite to disrupt the integrity of the systems on the spacecraft.

3. Methodology

This section discusses the research methodology, explains the motivation and approach, describes the system under test, and details the operationally representative test environment.

3.1 Problem Statement

Cyber attacks can alter the confidentiality, integrity, and availability of a system. Cyber threats to space systems are persistent, with increasingly severe effects against military, commercial and academic institutions. Satellite development teams that delay adding cybersecurity defenses until after launch incur unintended risk side-effects that may cost more in the long run. It is imperative to conduct cybersecurity assessments on space systems to mitigate potential cybersecurity vulnerabilities during the design and development. Cybersecurity assessments ensure that the systems released for production uphold a secure cybersecurity posture.

3.2 Motivation

There is a gap in testing the space vehicle of a satellite system. Often, it is due to budget constraints, lack of suitable simulation, or prioritization to allocate resources towards cyber testing. Often, system stakeholders conduct cybersecurity vulnerability assessments on satellite systems' ground stations because they resemble traditional information technology network structures in non-space systems. The immature standard on conducting cybersecurity assessments for the space vehicle of satellite systems warrants attention, especially in small satellites where the technology is more readily accessible.

Discussing cyber vulnerabilities on satellite systems is a sensitive topic. The studies and research done on military or commercial satellite systems are often classified or kept close hold. The minimal amount of public and unclassified research in cybersecurity testing satellites affirms the need for research on this topic. Research is particularly scarce when focused on hacking and causing adverse effects to satellite systems through the satellite communication links.

3.3 Approach

This research produces an assessment of cybersecurity vulnerabilities and attack vectors that threaten a space vehicle. The approach sets up an operationally representative test environment for the system under test. The data flow in this experiment must parallel an environment operating in the real world. The best approach to create this test environment is to incorporate a FlatSat. A FlatSat is a large motherboard that includes the satellite components installed and connected as if it was the actual satellite. The environment connects to a computer that hosts virtual machines (VM) to simulate the ground and flight software in the test setup. The FlatSat used in this experiment consists of two components from the satellite: an electrical power supply and a telemetry, tracking, and commanding component. The VM emulates the rest of the components in software. Connections between the computer, SDR, and the FlatSat create a test environment that accurately simulates the uplink and downlink between the satellite and the ground station.

3.4 System Under Test

Small satellites, characterized by their low cost, and small size, enable organizations to conduct their mission without the complexity of a large satellite. Their size and weight make it more affordable to launch into orbit. Due to low costs, it creates the ability to reach space widely accessible to organizations and countries when they previously could not. CubeSats are small satellites made for research. A 10 cm x 10 cm x 10 cm cube makes up one unit or 1U. CubeSats have shorter development times enabling launch within a couple of years, whereas conventional satellites could take 5 to 15 years. Low cost and short development times allow CubeSats to launch into space every 2 to 4 years to start new research or even update technology (Space, 2019).

This research paper will refer to the 6U CubeSat used in this experiment as "SmallSat". The SmallSat uses open-source flight and ground software, cFS and COSMOS, respectively. This particular SmallSat is a research satellite with a primary objective to fly the software bus and communicate to the ground station.

3.4.1 COSMOS

COSMOS is the ground software or front-end application used by operators in the satellite operations center. It is an open-source software developed by Ball Aerospace that provides a user interface and functionality to command and control space systems (Ball Aerospace, 2020). For the SmallSat, operators interface COSMOS to send commands to the spacecraft.

3.4.2 Titan/Blockbuilder

Titan is the software interface to the Universal Software Radio Peripheral (USRP) N210, designed to communicate to the Cadet Plus. Blockbuilder is the software tool that provides signal processing blocks to implement to USRP N210. For this paper, Titan and Blockbuilder are interchangeable terms. The result of Titan is a waveform and message format that the Cadet Plus accepts. This experiment uses Titan unaltered for the SmallSat. Figure 2 is a snippet of Titan showing the global variables declared for the uplink. This XML file is the configuration sent to the USRP to process the waveform to send to the Cadet Plus.

```
▼ <BLUEPRINT>
  <GLOBAL name="g_CommandPort" val="30555"/>
  <GLOBAL name="g_UsrpIp" val="192.168.10.2"/>
  <GLOBAL name="g_CadetAddress" val="0x6D"/>
  <GLOBAL name="g_TxCenterFrequency" val="449.775e6"/>
  <GLOBAL name="g_LogLevel" val="info"/>
  <GLOBAL name="g_TxBitRate" val="9600"/>
  <GLOBAL name="g_TxSampleRate" val="2000000"/>
  <GLOBAL name="g_TxGain" val="0"/>
  <GLOBAL name="g_BurstCount" val="16"/>
  <GLOBAL name="g_BurstDelayMs" val="150"/>
▼ <CHAIN name="Uplink">
```

Figure 2: Titan XML Configuration

3.4.3 Cadet Plus Radio

The Cadet Plus radio is a split band, full-duplex, store, and forward radio (Space Dynamics Laboratory, 2017). It is the physical component that operates the telemetry, tracking, and commanding of the satellite. The Cadet Plus is the uplink and downlink interface for the spacecraft. This experiment focuses on the ultra-high frequency uplink configured with a frequency range of 449.75 Mhz to 450.25Mhz, modulated at 9600 bps, as defined in the Interface Control Document (ICD).

3.4.4 Core Flight System

NASA's cFS is an open-source platform consisting of reusable software frameworks and software applications (Wilmot and Kane, 2021). In the SmallSat, the Command & Data Handler (C&DH) is the physical board that hosts cFS. For the experiment, cFS is emulated on the Titan VM.

3.5 Test Environment

The virtual test environment simulates the data being transmitted and received by the ground station and spacecraft. Hardware equipment required for this setup includes a desktop to host a VM, USRP N210, Cadet Plus radio, power supply, and cables for connection. Software required for this setup includes an Ubuntu Titan VM,

COSMOS, cFS, Blockbuilder, GHex, and Wireshark. USRP N210 was the SDR used in this experiment because it is the exact SDR model used for the system under test. This avoids disparity and variance. Figure 3 depicts the test setup architecture for the scope of this experiment.

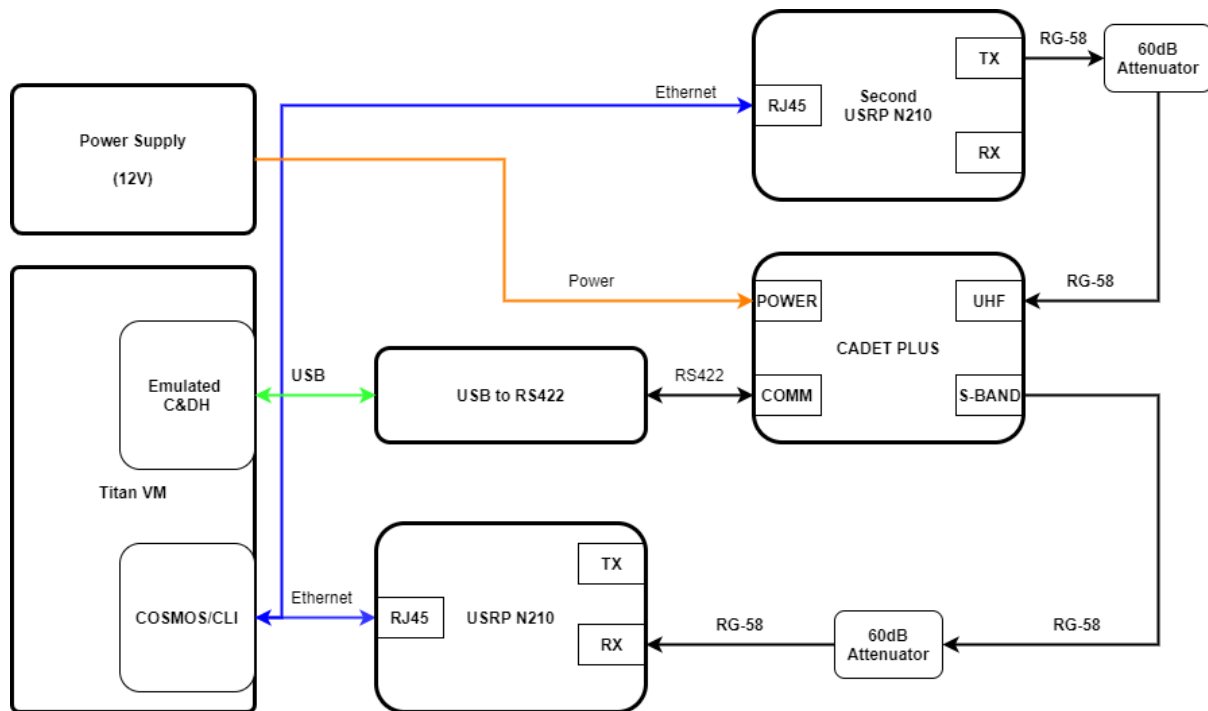


Figure 3: Test Setup Architecture

SDRs connect to the Titan VM via Ethernet cables. SDRs forwards the modulated data to the Cadet Plus radio, thereby simulating the data passing to and from the ground station and the spacecraft. Flight software runs on a VM in an emulated C&DH rather than a physical C&DH board. Cadet Plus radio then forwards validated data to the emulated C&DH, where the rest of the data processing occurs. Figure 4 depicts the actual test setup in the lab.

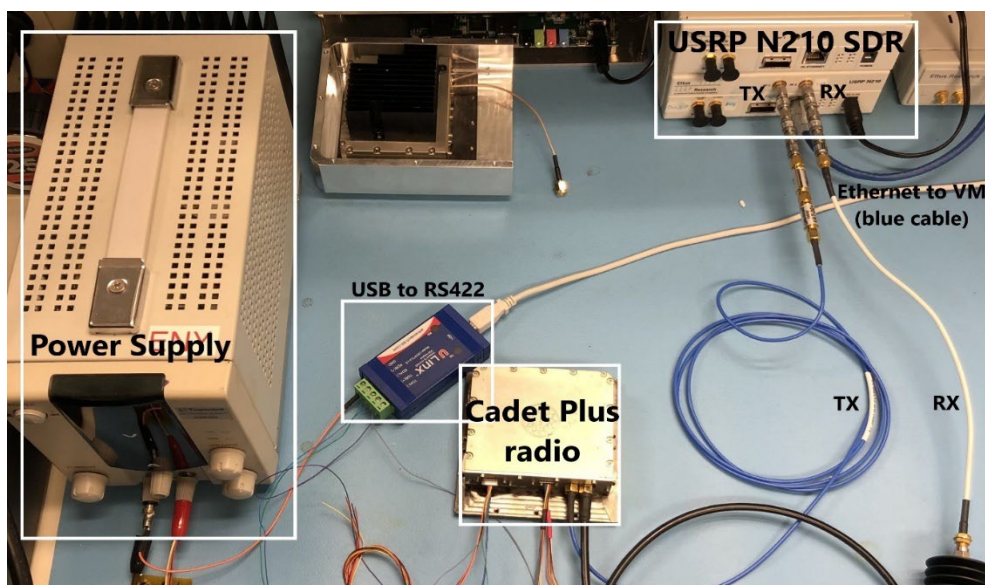


Figure 4: Image of Test Setup in Lab

4. Test Scenario and Results

This section discusses the test scenario of interest and the experimental methodology.

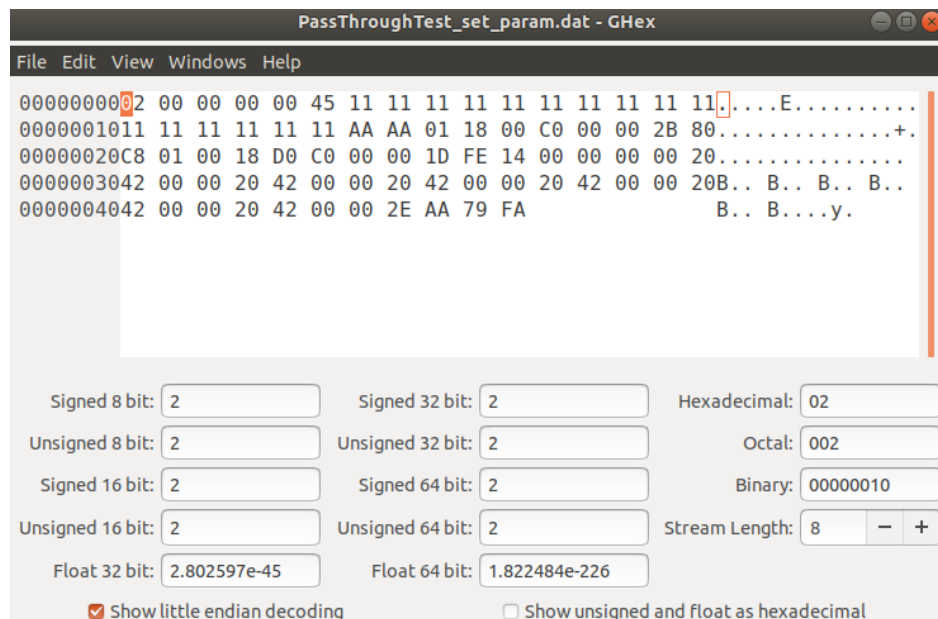


Figure 3: Edit Hexadecimal Data in GHex

The Cadet Plus ICD was used to verify the message format required to send to the USRP. Every byte must match the appropriate field, as portrayed in Table 1.

Table 1: General Cadet Plus Message Format (Space Dynamics Laboratory, 2017)

Bytes	Field
0-5	Space Packet Header
6	Flags (CRC, Quiet, Reserved)
7	Message ID
8-9	Dialog ID
10+N	Payload Data
10+N+1	CRC

The script portrayed in Figure 7 sends data in a .dat file specified with the source and destination port.

```
while [[ $(date -u +%s) -le $endtime ]]
do
    echo $(date +%s) $(cat PassThroughTest_set_param.dat)
    cat PassThroughTest_set_param.dat | nc -p 1337 -q $waitTime $addr 30555
    sleep 5s
done
```

Figure 4: Script to Run Command

Sending the command with a script replaces the need to send the command from COSMOS. Any individual with access to a CLI and a compatible RF software-defined radio can send commands to the spacecraft.

An additional *printf* statement that prints the hexadecimal data from the altered command verifies whether the altered command was successful. If the data passes the checks on the Cadet Plus, it is sent to the software bus. Once the data is on the software bus, cFS will display the *printf* statements, shown in Figure 8. The output from these *printf* statements verifies a successful command execution.

```
// Check for a completed message and send it to SB
if(sequence->ActualBytes == sequence->ExpectedBytes) {

    printf("Message in hex: \n\t");
    for(uint16 i = 0; i < sequence->ActualBytes; i++) {
        printf("%02X ", sequence->Buffer[i] );
        if(i%8 == 7) {
            printf("\n\t");
        }
    }
    printf("\n");

    CI_SendMsg(sequence->RadioID, &(sequence->Buffer));

    // Clear the sequence
    memset(sequence, 0, sizeof(CI_MessageSequence_t));
} else if(sequence->ActualBytes > sequence->ExpectedBytes) {
    CFE_EVS_SendEvents(CI_INVALID_MSG_SEQ_ERR_EID,
                       CFE_EVS_ERROR,
                       "CI Message sequence is invalid  

                       (More bytes received than expected)")
```

Figure 5: *Printf* Statements on Microsoft Visual Studio for Verification

The last step to the attack implementation is to execute the command as the insider. To begin test execution, the insider sends the command via CLI to execute the script. Figure 9 depicts the script from a terminal in the virtual test environment during test execution.

```
titan@titan-VirtualBox:~/Documents/Banks$ ./sendPassThrough1_2.sh  
./sendPassThrough1_2.sh: line 27: warning: command substitution: ignored null by  
te in input  
1632249430 [REDACTED] + [REDACTED] [REDACTED] B B B B B .oy
```

Figure 6: Execute Script from Terminal

4.3 Results

After executing the script, the flight software displayed the altered command in hexadecimal and prompted the following messages, shown in Figure 10: "Message sequence successfully completed and sent to the SB" and "MAI401: Set Parameter Command Success". These messages confirms that the message sent from the script was successful and spacecraft executed the command as written.

```

HS: The CPU is alive.
Message in hex:
    18 D0 C0 00 00 1D FE 14
    00 00 00 00 20 42 00 00
    20 42 00 00 20 42 00 00
    20 42 00 00 20 42 00 00
    20 42 00 00
EVS Port1 42/1/CI 23: Message sequence successfully completed and sent to the SB
EVS Port1 42/1/CI 20: CI: Message was sent to SB
◆◆ B B B B B B
 EVS Port1 42/1/MAI401 32: MAI401: Set Parameter Command Success
HS: The CPU is alive.

```

Figure 7: cFS Shows Message in Hex *Printf* Statement Additions

The Cadet Plus radio accepted the command and pushed it onto the software bus. The altered command successfully executed and forced the satellite to change the *sunVec_b* sensor values to detect its exposure as "full sun" due to the altered *CssGain* value. Figure 11 highlights the attack path in red over a data flow diagram of the system between ground station and spacecraft. It depicts that the command was able to reach to the MAI401 can cause impact.

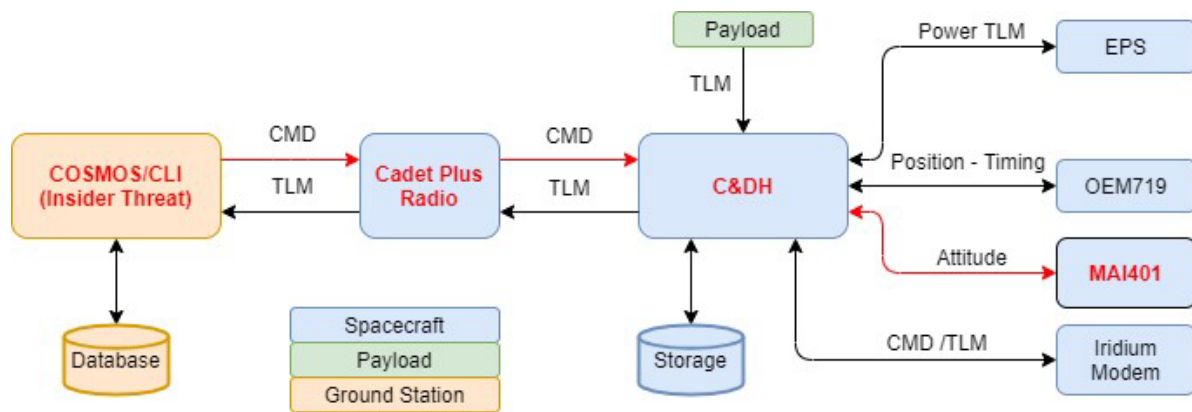


Figure 8: Insider Attack Path

The experimental scenario successfully impacted the satellite's integrity by changing the *CssGain* value, tricking the satellite sensor into believing it is extremely high or at "full sun" exposure. Using the "SET_PARAMETER" command to change the *CssGain* value is just one of many options to alter commands for malicious intent. The ADACS has many more commands, from changing *CssBias* that affects the sun sensors to setting the Cadet Plus to test mode and locking the configurations.

Capturing the entire data with Wireshark was a practical approach to generating a file containing the raw hexadecimal of the altered command. Although this attack may seem unlikely outside of an insider threat scenario, a determined adversary can learn the correct message format and CRC calculation given enough time and resource. The fact that unauthenticated commands can be accepted by the space system without COSMOS proves that it is possible to spoof uplink commands from other untrusted sources. The primary factor is that the raw data format must meet the configurations accepted by the Cadet Plus radio.

5. Conclusion and Future Work

In conclusion, the initial experiment demonstrates the capability to send an altered command with malicious intent to the satellite without using COSMOS to send the command. The attack successfully changed the value of *CssGain*, unjustly exposing the sensor to "full sun." The results of this experiment can be replicated for commands only requiring one uplink packet without requiring an acknowledgment.

Software configurations for the USRP N210, remained unaltered for this experiment. It ensures that the correct waveform would reach the Cadet Plus. Since the Titan and USRP N210 are software-defined, an adversary can use open-source software alternatives to Titan, such as GNU Radio Companion, to accomplish the same objective. Adversaries can configure their SDR settings to meet the requirements for Cadet Plus. Future work will investigate sending altered commands external to the ground segment to signify an outsider threat. It requires a well-configured SDR that interfaces with a separate USRP N210 to send the command to Cadet Plus. Additional future work includes characterizing the risk of satellite commands and associating its risk to specific subsystems on the spacecraft.

The space enterprise cannot neglect the legitimate concerns of an insider threat. However, if this cyber-attack demonstrated in this paper evolves into an outsider threat, the impact on space systems can be detrimental. The cyber-attack vector like the ones described in this paper may apply to small satellites used in universities and organizations that run NASA's cFS. The ability to identify, defend, and mitigate these cyber-attacks will position the global space enterprise in a better cybersecurity posture from potential attacks and disruption.

Disclaimer

The views expressed are those of the author and do not reflect the official policy or position of the US Air Force, US Space Force, Department of Defense, or the US Government.

References

- Adcole Marland Aerospace, Inc (2020). MAI-401 Mini ADACS. pp.36, 38, 49, 70.
- Ball Aerospace (2020). COSMOS. [online] Ball. Available at: <https://www.ball.com/aerospace/programs/cosmos>.
- Book, G., 2015. Security Threats against Space Missions. Informational Report.

- Falco, G., 2019. Cybersecurity principles for space systems. *Journal of Aerospace Information Systems*, 16(2), pp.61-70.
- Gupta, A. (2017). *The IoT Hacker's Handbook*. Apress, pp.223–263.
- Hung, P.D. and Vinh, B.T. (2019). Vulnerabilities in IoT devices with software-defined radio. In: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). IEEE, pp.664–668.
- Jameel, S. (2020). Security Through Obscurity: Valid Security Layer or Bad Idea? [online] cgscomputer.com. Available at: <https://cgscomputer.com/cyber-security-through-obscurity/>.
- Lukin, K. and Haselberger, M. (2020). Hacking Satellites With Software Defined Radio. In: 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). IEEE, pp.1–6.
- Manulis, M., Bridges, C.P., Harrison, R., Sekar, V. and Davis, A., 2021. Cyber security in new space. *International Journal of Information Security*, 20(3), pp.287-311.
- Nussbaum, B. and Berg, G. (2020). Cybersecurity implications of commercial off the shelf (COTS) equipment in space infrastructure. *Space infrastructures: From risk to resilience governance*, pp.91–99.
- Space Dynamics Laboratory (2017). Cadet Plus Radio Interface Control Document. Space Dynamics Laboratory (SDL), pp.14–15.
- Space, A. (2019). 10 Advantages of CubeSats vs. Conventional Satellites. [online] Alén Space. Available at: <https://info.alen.space/advantages-of-cubesats-vs-conventional-satellites>.
- Tucker, P. (2019). The NSA Is Studying Satellite Hacking. [online] Defense One. Available at: <https://www.defenseone.com/technology/2019/09/nsa-studying-satellite-hacking/160009/>.
- Vessels, L., Heffner, K. and Johnson, D., 2019, July. Cybersecurity Risk Assessment for Space Systems. In 2019 IEEE Space Computing Conference (SCC) (pp. 11-19). IEEE.
- Waterman, S. (2019). Satellite Providers Stymied by Lack of Cyber Standards - Via Satellite -. [online] Via Satellite. Available at: <https://www.satellitetoday.com/cybersecurity/2019/11/14/satellite-providers-stymied-by-lack-of-cyber-standards/>.
- Wheeler, W.A., Cohen, N., Betser, J., Meyers, C., Snively, W., Chaki, S., Riley, M. and Runyon, B. (2017). Cyber Resilient Flight Software for Spacecraft. In: *AIAA SPACE and Astronautics Forum and Exposition*. p.5305.
- Wilmot, J. and Kane, L. (2021). Core Flight System. [online] cfs.gsfc.nasa.gov. Available at: <https://cfs.gsfc.nasa.gov/Introduction.html>.

Improving Hardware Security on Talos II Architecture Through Boot Image Encryption

Calvin Muramoto, Stephen Dunlap and Scott Graham

Air Force Institute of Technology, USA

Calvin.Muramoto@afit.edu

stephen.dunlap.ctr@afit.edu

Scott.Graham@afit.edu

Abstract: The OpenPOWER Foundation is an organization that promotes open-source high-performance hardware like the POWER9. OpenBMC is an OpenPower project that strives to produce an open-source firmware stack for Baseboard Management Controllers (BMCs). If hardware falls into the hands of competitors or bad actors, reverse engineering methods can be used to leak or manipulate sensitive information from the boot sequence. This represents a security concern because the root of trust can be invalidated. For example, since the Initial Program Load (IPL) data is frequently not encrypted and is sent over the Low Pin Count (LPC) bus, it is possible to intercept and conduct man-in-the-middle attacks to modify the boot process. The boot image flash chip could also be removed from the Talos II motherboard and examined by competing server architecture manufacturers to reveal detailed boot information. Firmware that developers deem to contain sensitive code or perform innovative operations needs to be protected before being flashed onto the boot image chip. This paper demonstrates a method to encrypt sections of the boot image by encrypting a section of the image before flashing it onto the Talos II. The encrypted image will be decrypted during the boot sequence in the Level 3 cache of the POWER9, proving that it is possible to prevent adversaries from interfering with the IPL flow or obtaining details on firmware from the flash chip. This paper presents a novel method to improve the security of the boot image on Talos II architecture by encrypting the boot firmware image and decrypting it during the boot process. The proof of concept was executed on a Raptor Engineering Talos II system running a POWER9 processor with OpenBMC firmware on the ASPEED AST2500 BMC. This research claims that this unique method increases boot time security through firmware without altering hardware.

Keywords: Secure Boot, Hardware Security, Firmware Encryption, PNOR, Boot Image, Hostboot

1. Introduction

As the complexity of cyberattacks increase and evolve, the demand for increased security in computer systems also rises. Hardware attacks have become a concern in recent years because physical device hardware is not normally designed to be resistant to the increasing number of attack vectors. To counter this threat, trusted computing has become a focus for computing system development with security enhancements, like secure boot (Mitchell, 2005). Although secure boot can prevent adversaries from loading malicious firmware on a system, no known system encrypts the existing boot firmware.

Protecting boot firmware data from being leaked allows developers to conceal code that performs innovative functionality to the system. Hiding this low-level firmware also increases the difficulty of discovering zero-day attacks or vulnerabilities in the code. Tamper detection methods should be located in source code in the boot sequence to prevent intruders before the operating systems starts. Leaking the source code on intrusion prevention systems would allow bad actors to avoid the tamper checks and gain access to the system. Boot firmware is often left unencrypted in memory leaving the system open to hardware attacks like bus traffic recording or chip substitution attacks. Encryption adds an additional layer of protection from adversaries looking to reverse engineer firmware.

Although this proof of concept is developed specifically for the Talos II, it should be possible to implement boot firmware encryption on common systems with Intel and AMD processors. As long as the system uses a similar boot flow structure as OpenPOWER, encrypting sections of the boot firmware should be similar.

This paper demonstrates how encrypting of a section of the boot image is possible on an OpenPOWER workstation. As a proof-of-concept, this effort used a simple exclusive OR function as a proxy for encryption. Success will be measured by implementing the simple encryption function, flashing the modified image onto the Talos II, and verifying that the system can successfully boot consistently. This would protect the boot data from outsiders by only updating the firmware.

This paper is split into 6 sections to cover the research effort. Section 2 introduces the background information required to understand the nature of boot image encryption. In section 3, the methodology for this proof of concept is covered along with a discussion on feasibility. The experimental analysis of the methodology along with definition of success is contained in Section 4. The impact of this research is covered in Section 5 and the conclusion and discussion on future work is covered in section 6.

2. Background

2.1 Talos II Architecture

Talos II is a secure workstation designed by Raptor Computing Systems. It makes use of dual POWER9 CPUs, PCIe 4.0, and DDR 4.0. Most importantly, the Talos II also supports OpenBMC firmware, allowing developers to utilize open-source Baseboard Management Controller (BMC) firmware. This allows us to modify the booting process through firmware in this study. The BMC is a functional system processor that can monitor and control the system that it is implemented on.

Initial Program Load (IPL) refers to the steps that OpenPOWER systems take from power on to the start of the hypervisor. During IPL, the Talos II utilizes several interfaces to transfer information between the PNOR flash chip, Serial electrically erasable programmable read-only memory (SEEPR0M), Self-boot engine (SBE), and POWER9. The most important interfaces are the Low pin count (LPC) bus, pervasive interconnect bus, and serial communication. The LPC bus loads the PNOR flash image to the POWER9 and can perform memory, I/O, direct memory access, and bus master cycles. A trusted platform module on the Talos II uses the LPC to communicate with the SBE and POWER9. The pervasive interconnect bus connects and facilitates transactions between multiple pervasive interconnect bus master and slave units inside the POWER9 chip. It is mainly used by the SBE at the very beginning of the IPL for memory I/O. Finally, serial communication is a mechanism in the POWER9 chip pervasive logic that gives access to pervasive registers when the clocks are running. It is a specialized interface into the chip pervasive logic that allows specific latches to be updated while functionally in use.

An important consideration taken into this research was the size available on the PNOR flash chip, SEEPR0M, and POWER9 Level 3 (L3) cache. Adding bloated code or having excessive memory usage during the IPL could cause a buffer overflow and crash the boot sequence. The PNOR flash chip is 4MB, the POWER9 SEEPR0M is 64 KB, and POWER9 L3 cache is 10 MB. When allocating space for temporary data in the XOR function, we had to ensure that the space did not exceed the L3 cache size. Although the L3 cache has a memory space limitation, the cache provides fast data transfer speeds to the POWER9 CPU cores which helps reduce memory I/O instruction execution time. We also had to ensure that the compiled XOR code did not overflow the memory allocated for the Hostboot firmware on the PNOR.

2.2 PNOR Image Structure

The PNOR Image contains all the firmware needed to boot the Talos II. It is split into logical partitions, with each section containing a firmware module. Figure 1 shows a general overview of the architecture of the boot image used in OpenPOWER systems. The Talos II version of the PNOR is 4 MB and contains 31 sections with specific functions like Secure Boot, DIMM SPD Cache, and Module VPD Cache (Raptor CS, 2019). The Table of Contents (TOC) is located at the start and end of the image and contains important data like the name of the partition, physical offset, and physical size. It is queried often during IPL to inform the firmware of the information needed to load the following partitions. Before loading each section in the IPL, the SBE or POWER9 will look up the address and size of each partition to load into the L3 cache or DRAM. The sections relevant to this research are the Hostboot Base (HBB), Hostboot Extended (HBI), and OPAL sections.

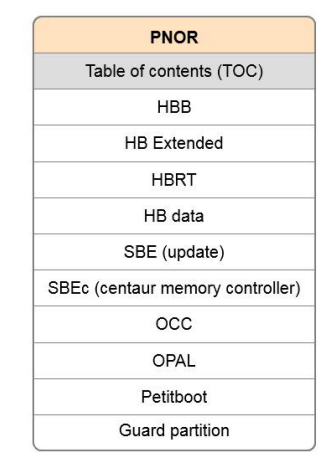


Figure 1: PNOR structure (IBM, 2019)

2.3 Initial Program Flow

IPL is a term that covers the boot process from power on to running the hypervisor. The Talos II follows the boot flow specified in Fig 2. Once the system is powered on, the OTPROM is loaded from the CPU die and executed on the SBE. The SBE is a specialized small auxiliary microprocessor that initializes the processor chip (IBM, 2019). The OTPROM is a section of SBE firmware that is permanently loaded through eFuses on the POWER9 silicon and loads the SEEPROM into the SBE pervasive interconnect memory (Raptor CS, 2021). The pervasive interconnect bus memory is a 64 KB SRAM attached to the pervasive interconnect bus and used by the SBE as its memory.

The SEEPROM firmware on the SBE core initializes a designated CPU core, allows access to the PNOR flash memory, and then loads the HBB from SEEPROM. SEEPROM is an EEPROM memory device that can only be read but can also be reprogrammed through the IPL process (Geissler, 2015). The stages that follow are all included in the PNOR and are covered in the following sections.

The PNOR flash chip is a NOR memory device where all firmware is stored for the latter portion of the IPL. It contains Hostboot, OPAL, the OCC, and Bootloader firmware. It is connected to the master or alternate master processor through the SPI bus. The image loaded into the PNOR consists of several sections of data that correspond to stages in the initial program load. This paper focuses on the Hostboot Extended (HBI) image and the payload.

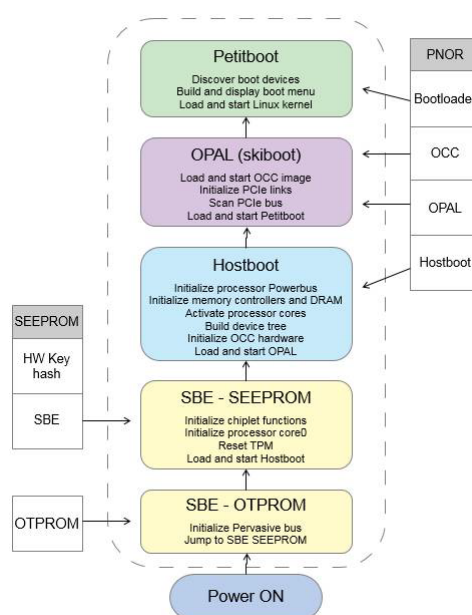


Figure 2: IPL Flow (IBM, 2019)

2.4 Hostboot

The Hostboot stage of the boot process is extremely important as it initializes a series of capabilities that are needed for later boot stages. It is loaded and started in ISTEP5, and covers ISTEPs 6 through 21. Throughout the course of these ISTEPs, the processor bus, memory controllers, DRAM, and PSI/Node are all initialized. The most important partitions are the HBBL, HBB, and HBI. Hostboot has its own cache-contained operating system that separates a user and kernel space along with task management for HBI (Jeffery, 2018). Since the memory requirements are larger than the 10 MB available on the L3 cache, a virtual memory and filesystem layer is used. This requires demand-paging to bring code and data in from the PNOR when needed (Jeffery, 2018).

The HBBL firmware is stored and run from SEEPROM but is also part of the PNOR image. This is because the HBBL on SEEPROM can be updated through a successful boot process. HBBL handles locating the HBB partition, ECC verification, loading into the L3 cache, and execution on the first core. This firmware is the first code that will run on the main CPU cores after power on. Although the HBBL source code is located in the Hostboot repository, it is stored on the SEEPROM SBE image where the secure boot root of trust hash is located. This is because the HBBL is responsible for cryptographically verifying the integrity of the HBB and HBI when trusted boot is enabled. The switch from HBBL to HBB is handled by the *bl_start.S* code.

HBB contains the hostboot kernel and handles key low-level modules like serial communication, LPC, PNOR access, and Virtual File System (VFS). HBB takes care of the required services to read and write to the PNOR. It also functions as the foundation for the rest of the HBI startup.

Execution of the HBI is done through the VFS. The HBI phase of the boot process consists of a list of ISTEPs and services that need to be executed. The *init-service* module runs through the list of services and sends a start message for each service. The Virtual File System Resource Provider (VFSRP) utilizes a watcher that handles each message by requesting the relevant memory addresses and responding. HBI is also responsible for chain-loading the subsequent firmware payload in ISTEP 20 and 21. When complete, Hostboot passes control of the CPU to the payload firmware and Skiboot is run.

2.5 Payload

On all OpenPOWER systems, the firmware payload is Skiboot. Skiboot handles the initialization of PCIE, device trees, real-time clock, NVlink, and sensors. These sensors are required when Skiboot loads the OCC firmware and begins its execution. The OCC manages CPU temperatures, power measurement, and other chassis-specific data. Finally, the OpenPOWER Abstraction Layer (OPAL) is started for OS runtime services. OPAL resides in RAM after OS boot. Runtime API implemented by firmware is called into by the host operating system to invoke platform-specific functionality. Skiboot then chain loads Skiroot and Petitboot, which act as bootloaders for Linux.

3. Methodology

The methodology for this proof of concept is complex due to the multiple stages required to make firmware changes. Before the PNOR is flashed onto the Talos II, a section of memory needs to be identified for encryption. The decryption function is then programmed into the firmware that loads the section of memory. For example, since the Hostboot stage chain-loads the payload stage, the payload would be encrypted and the decryption function would be located in Hostboot. The PNOR is then compiled and the address space identified earlier is encrypted. The PNOR with the encrypted section and decryption function is flashed onto the Talos II. Multiple successful boot sequences would be proof that the implementation was successful.

3.1 Proof of Concept Development Discussion

After analyzing the IPL firmware, several points in the code were identified where the decryption code could execute successfully. The first possibility was the Hostboot Bootloader (HBBL) which was located in *bootloader.C*. A function called *handleMMIO()* is used to copy the HBB code from the PNOR to a working location. The HBBL code is executed near the start of the boot process, so machine code is used to communicate with the LPC to transfer data. This also means that the console output (print out and log commands) would not be accessible, making it difficult to verify the correct address and size of the partition for encryption.

Another problem was the risk in altering the firmware before HBB in the IPL because the SEEPROM uses two sides to save low-level firmware while keeping a backup. When the Talos II boots up with new firmware, the new HBBL is saved to side 0. If the system successfully runs through the HBBL once, the system saves the new

firmware to side 1. A critical problem may occur if the firmware successfully boots once but fails in subsequent runs, resulting in permanent failure to boot, i.e., the POWER9 could get “bricked”.

The second location in code possible was for the HBI. Since HBB handles the modules for starting the HBI, we were able to narrow down the location to the *vfsrp.C*, which is the virtual file system resource provider. HBI is executed through ISTEPs and a module list which each have their own services to initialize. HBB reads through the module list and sends messages to the POWER9 to load in each task and initialize them. Each task sends a MSG_MM_RP_READ, which executes a *memcpy()* which is used to load the relevant data into the local execution space. The problem with using VFSRP is that a maximum of 68 tasks may be started during the HBI phase. This means that 68 addresses and address spaces must be managed during the encryption phase making this implementation extremely complicated.

Finally, the last point in code is located at the end of the HBI in *call_host_load_payload.C* in ISTEP 20. This code is responsible for calling a function, *load_pnor_section()*, which is only run one time to load the payload from the PNOR into local memory space. The only possible problem with working in this code would be the XZ compression. Since the payload section is nearly 16 MB but is only allocated 1 MB on the PNOR image, it must be compressed when stored on the flash chip. The decompression code handles the *memcpy()* from PNOR but raised concerns with read and write operations with regard to the LPC address space.

The plan was to allocate a space to store the encrypted payload, as shown on line 270 in Figure 3, and copy the data from the PNOR to the local memory space. The encrypted payload was loaded in 4 KB chunks, as shown in lines 277 to 284 in Figure 4. After all the encrypted data was stored in the *temp_buf*, the data was XORed in place on lines 289 to 292 in Fig 4. The XORed data was then passed into the XZ decompression function. The *display* function from the console was used to display important data to the BMC console client and helped determine if the code failed at a specific section.

```
250 // Display a banner on the status of PNOR compression
251 console::displayf(NULL, "call_host_load_payload.C: PNOR is XZ compressed\n");
252 console::flush();
253
254 struct xz_buf b;
255 struct xz_dec *s;
256 enum xz_ret ret;
257
258 xz_crc32_init();
259 s = xz_dec_init(XZ_SINGLE, 0);
260 if(s == NULL)
261 {
262     TRACFCOMP(ISTEPS_TRACE::g_trac_isteps_trace, ERR_MRK
263               "load_pnor_section: XZ Embedded Initialization failed");
264     return err;
265 }
266
267 console::displayf(NULL, "call_host_load_payload.C: Starting malloc\n");
268 console::flush();
269
270 uint8_t* temp_buf = (uint8_t*) malloc(originalPayloadSize);
271
272 if (! temp_buf) {
273     console::displayf(NULL, "call_host_load_payload.C: Malloc failed\n");
274     console::flush();
275 }
```

Figure 1: XOR code showing the XZ compression code and the allocated space for decryption.

```

277 const uint32_t BLOCK_SIZE = 4096;
278 for ( uint32_t i = 0; i < originalPayloadSize; i += BLOCK_SIZE )
279 {
280     memcpy( reinterpret_cast<void*>(
281         reinterpret_cast<uint64_t>(temp_buf) + i ),
282         reinterpret_cast<void*>( pnorSectionInfo.vaddr + i ),
283         std::min( originalPayloadSize - i, BLOCK_SIZE ) );
284 }
285
286 CONSOLE::displayf( NULL, "call_host_load_payload.C: Done copying to local buffer\n");
287 CONSOLE::flush();
288
289 for ( uint64_t i = 0; i < originalPayloadSize; i++ )
290 {
291     temp_buf[i] ^= 0x55;
292 }
293
294 for (uint32_t i = 0; i < 16; i++){
295     CONSOLE::displayf( NULL, "Header %d = %02x\n", i, temp_buf[i]);
296     CONSOLE::flush();
297 }
298 CONSOLE::displayf( NULL, "call_host_load_payload.C: Done XORing\n");
299 CONSOLE::flush();

```

Figure 2: XOR code showing the memcpy() into the allocated space to be XORed with the key.

3.2 Assumptions and Limitations

Since this research was a proof of concept to show that the PNOR image could be encrypted, a simple XOR was used to ‘encrypt’ the boot firmware. This is a limitation in this proof of concept because a simple XOR is not secure. A key of “0x55” was used when XORing the payload because it should flip every single bit in the section, making it easy to verify in a hex editor. For the purposes of this demonstration, a simple XOR is representative of an encryption scheme.

This implementation is also only executed on the payload of the boot sequence, meaning that only the payload would be secure. The decryption code in Hostboot is also in plaintext, along with the key used to XOR the partition. These limitations will be addressed in future research.

This implementation requires a Talos II system running a single POWER9 without secure boot. It also requires the PNOR image to be compiled on a separate system and XORed before being flashed onto the system.

Finally, the key of “0x55” was stored unprotected in the Hostboot firmware; mechanisms to properly secure such a key will be explored in future research.

4. Experimental Analysis

There was a problem when initially working with *call_host_load_payload.C* because the payload in the Talos II PNOR is XZ compressed. The memcpy() responsible for loading the compressed payload exists in the XZ decompression code, adding a layer of complication. In this implementation, we planned to allocate a space in memory to manually store the PNOR image, decrypt the compressed image, and pass the decrypted data into the decompression function.

A few attempts were made to boot the Talos II with an encrypted payload PNOR successfully. Initially, the XOR decryption function was used directly on the data pointed to by the *pnorSectionInfo.vaddr*. We quickly discovered that the virtual address specified for each PNOR section is memory mapped to an address space on the LPC bus so the system did not boot successfully. This is a problem because we were attempting to XOR in place, and this would attempt to write over the PNOR section over the LPC bus. We resorted to an implementation with a higher cost to memory by allocating a 1 MB block that would copy memory from PNOR and XOR the data in place. This memory space was then passed into the decompression function and led to a successful boot.

After successfully running all the steps, a standard boot operation was conducted as a test. This was done to verify that the XOR function did not interfere with OPAL and the following boot process. Figure 5 shows the OBMC console client logs where the print statements in Figures 3 and 4 are shown. The figure also shows ISTEP 21.3 on line 98 and the following log on line 99 show that OPAL Skiboot was able to start, meaning that the XOR code ran successfully. After running the boot sequence several times and cycling the power, we verified that the Talos II was able to boot Linux without issue.

```

69 33.84818|ISTEP 20. 1 - host_load_payload
70 33.87024|call_host_load_payload.C: Starting load_pnor_section()
71 33.87144|call_host_load_payload.C: PNOR is XZ compressed
72 33.87391|call_host_load_payload.C: Starting malloc
73 34.53205|call_host_load_payload.C: Done copying to local buffer
74 34.53735|Header 0 = fd
75 34.53027|Header 1 = 37
76 34.53836|Header 2 = 7a
77 34.53878|Header 3 = 58
78 34.53926|Header 4 = 5a
79 34.53970|Header 5 = 00
80 34.54022|Header 6 = 00
81 34.54065|Header 7 = 01
82 34.54111|Header 8 = 69
83 34.54157|Header 9 = 22
84 34.54202|Header 10 = de
85 34.54250|Header 11 = 36
86 34.54295|Header 12 = 02
87 34.54348|Header 13 = 00
88 34.54393|Header 14 = 21
89 34.54442|Header 15 = 01
90 34.54484|call_host_load_payload.C: Done XORing
91 34.54563|Running Decompression
92 34.66993|Freeing buffer
93 34.67050|call_host_load_payload.C: Ending load_pnor_section()
94 34.67252|ISTEP 20. 2 - host_load_hdat
95 35.55996|ISTEP 21. 1 - host_runtime_setup
96 41.29561|htmtgt|OCCs are now running in ACTIVE state
97 46.02538|ISTEP 21. 2 - host_verify_hdat
98 46.05073|ISTEP 21. 3 - host_start_payload
99 [ 46.323404859,5] OPAL skiboot-9858186 starting...

```

Figure 3: Boot log messages and successful OPAL start message

5. Research Impact

The successful implementation of the PNOR encryption methodology in this research shows that it is possible to add a layer of security over the IPL without changing the hardware. The encrypted boot instructions are decrypted in the POWER9 processor and stored in the L3 cache, making it extremely difficult to reveal the plaintext firmware. This also proves that it is possible to manipulate the boot data if the allocated space does not overflow memory. The ability to control boot data during the IPL shows that the capabilities of the POWER9 are only limited by the memory space available during the early stages of the boot process.

This approach is expected to be applicable on server architectures utilizing other processors, such as Intel and AMD. It should be possible if the boot structure follows a similar model to OpenPOWER's format. This would allow every consumer computer system to support encrypted boot firmware which would help protect against hardware attacks. The boot firmware encryption technique could also be applied to BIOS and firmware updates, permitting new boot firmware to be securely flashed onto a computer system without the risk of source code being leaked.

6. Conclusions and Future Work

Since this research was a proof of concept to show that the PNOR image could be encrypted, a simple XOR was used to 'encrypt' the boot firmware. The implementation of the XOR can be changed to a secure encryption like SPECK, AES or DES in the future. SPECK is a lightweight encryption protocol developed by NSA optimized for firmware implementation (Beaulieu, 2015).

This research also only focused on implementing the boot image encryption on the OPAL section of the PNOR. In the future, this could be expanded to multiple partitions based on how the firmware stage is loaded into memory. The multi-section encryption could be accomplished by programming the decryption method into each partition and encrypting the respective portions.

Future research will be to incorporate the PNOR encryption scheme with Secure Boot enabled. IBM's Secure boot allows for the verification of Hostboot and stores the source code on the SBE SEEPROM with the goal of ensuring that only legitimate firmware is loaded into the flash chip (Heller, 2019). This should be relatively simple as secure boot uses a chain of trust system where each firmware component of the boot process verifies the next section before being allowed to run.

Finally, key management would need to be solved as the current key exists unprotected in the HBI firmware. In future research, the key should be saved in SEEPROM with the verification and bootloader signatures.

Disclaimer: The views expressed in this paper are those of the authors, and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government. This document has been approved for public release; distribution unlimited, case # 88ABW-2021-0886.

References

- Beaulieu, R., Shors, D. and Smith, J. (2015) "Simon and Speck: Block Ciphers for the Internet of Things". National Security Agency. <https://csrc.nist.gov/csrc/media/events/lightweight-cryptography-workshop-2015/documents/papers/session1-shors-paper.pdf>.
- Geissler, A. (2015) "Hostboot POWER Systems Initialization Firmware". <https://github.com/open-power/hostboot>. Accessed 20 Sep 21.
- Jeffery, A. (2018) "General Architecture of Hostboot". <https://amboar.github.io/notes/2018/08/19/hostboot-architecture.html>. Accessed 20 Sep 21.
- Jeffery, A. (2018) "Hacking Hostboot". <https://amboar.github.io/notes/2018/08/17/hacking-hostboot.html>. Accessed 20 Sep 21.
- Heller, D and Sastry, N. (2019) "OpenPower secure and trusted boot, Part 2: Protecting system firmware with OpenPOWER secure boot". <https://developer.ibm.com/articles/protect-system-firmware-openpower/>. Accessed 20 Sep 21.
- IBM (2020) "Secure Initial Program Load (IPL) process". https://www.ibm.com/docs/en/power9/9009-42A?topic=9009-42A/p9ia9/p9ia9_secure_ipl_proc_concept.htm. Accessed 20 Sep 21.
- Microchip (2017) "AVR231: AES Bootloader". <http://ww1.microchip.com/downloads/en/AppNotes/00002462A.pdf>. Accessed 5 Dec 21.
- Mitchell, C. (2005) "Trusted Computing". The Institution of Engineering and Technology, London, United Kingdom. p. 1-6
- OpenPOWER (2021) "Hostboot". <https://github.com/open-power/hostboot>. Accessed 20 Sep 21.
- Raptor Computing Systems (2019) "index : pnor". https://git.raptorcs.com/git/pnor/tree/p9Layouts/defaultPnorLayout_64.xml. Accessed 20 Sep 21.
- Raptor Computing Systems Wiki (2021) "OpenPOWER Firmware". https://wiki.raptorcs.com/wiki/OpenPOWER_Firmware. Accessed 20 Sep 21.

Technical Analysis of Thanos Ransomware

Ikuomor Ogiriki, Christopher Beck and Vahid Heydari

Rowan University, Glassboro, NJ, USA

ogirik95@students.rowan.edu

beckch36@students.rowan.edu

heydari@rowan.edu

Abstract: Ransomware is a developing menace that encrypts users' files and holds the decryption key hostage until the victim pays a ransom. This particular class of malware has been in charge of extortion hundreds of millions of dollars every year. Adding to the problem, generating new variations is cheap. Therefore, new malware can detect antivirus and intrusion detection systems and evade them or manifest in ways to make themselves undetectable. We must first understand the characteristics and behavior of various varieties of ransomware to create and construct effective security mechanisms to combat them. This research presents a novel dynamic and behavioral analysis of a newly discovered ransomware called Thanos. It was founded in 2020 and is building up to be the leading malware used by low-to-medium-level attackers. It is part of a new ransomware class known as RaaS (Ransomware as a Service), where attackers can customize it for their desired target audience. So far, it is more prevalent in the middle east and North Africa and has over 130 unique samples already. As part of this investigation, the Thanos ransomware is carefully being analyzed. A testbed is created in the virtual artificial environment that mimics a regular operating system and identifies malware interactions with user data. Using this testbed, we can study how ransomware generally affects our system, how it spreads, and how it continually persists to access the user's information. We can design a new security mechanism to detect and mitigate Thanos and similar ransomware based on behavior examination results.

Keywords: Ransomware, malware detection, Virtual Machine, Threat, Thanos

1. Introduction

Since the advent of cryptos like Bitcoin, cybercrime has been on the rise that led to the development of a new type of malware called ransomware. Ransomware is a malicious software that enables attackers to access users' data and restrict them until a ransom is paid. Ransomware has grown to become the most financially rewarding form of malware for cyber attackers over the years. There are two major types of ransomware: 1) encrypting these data until a ransom is paid to decrypt (crypt) using a strong cryptography algorithm, 2) locking users out without tampering with their data until the ransom is paid (locker). Since the introduction of ransomware, the threat landscape has witnessed an increase of an annual rate of 267 percent. This research presents a novel dynamic and behavioral analysis of a newly discovered ransomware called Thanos.

It's essential to look at previous approaches that have been aimed at analysing and detecting ransomware. Previous research has attempted to identify and eliminate malware infections using the signature-based strategy. However, so far though, this has shown to be ineffective in properly identifying all types of malware. While signature-based solutions are rapid and efficient, they are also easily circumvented by very new or much older kinds of malware as discussed by MahdaviFar and Ghorbani (2019) and Zhang et al (2020). The behavior-based malware detection approach, on the other hand, has stronger resistance to this older virus but falls short since it is exceedingly time-consuming. We may conclude from these two distinct types of malware detection systems that, while both are effective solutions, they are not comprehensive enough to recognize malware.

As a solution, we propose a more hybrid system, utilizing both the signature and dynamic behavior-based detection techniques, with the added layer of the machine learning algorithm. This hybrid system would give a more robust answer to the malware detection challenge.

The primary purpose of this research is to present multiple potential approaches for training malware classifiers using machine learning. Another objective is to prevent ransomware from spreading laterally inside a network. To prevent the lateral movement of ransomware, we will use various techniques such as SDN, Moving Target Defense, and Zero Trust.

2. Related Work

This section includes a brief review of previous ransomware detection methods classified based on machine learning, file system/process monitoring, and network traffic.

2.1 Classifications Based on Machine Learning

2.1.1 DNAAct-Ran

According to Khan et al (2020), the traditional signature-based malware detection is no longer efficient in identifying ransomware. As a result, their answer presented a new and better way using the ground-breaking Digital DNA sequencing engine. This engine employs a machine-learning algorithm to classify ransomware based on its digital genome and phenotype to identify its numerous destructive functionalities appropriately. The suggested DNAAct Ran technique uses machine learning to determine if a software is a ransomware. This technique achieves this aim by first selecting the significant traits, generating digital sequences for those selected futures, and detecting the ransomware. The algorithms utilized are the Multi-Objective Grey Wolf Optimization (MOGWO, an extension of GWO by Mirjalili, Mirjalili and Lewis (2014)) and the Binary Cuckoo Search (BCS algorithm) by Yang and Suash (2009). The classifier's performance in comparison to other ML approaches is used to assess the accuracy of the DNAAct-capacity Ran's to identify ransomware. Some other active machine learning approaches, in addition to this, suggested one, include naive Bayes, decision stump, and the Adaboost classification algorithm.

2.1.2 Know Abnormal, Find Evil: Frequent Pattern Mining for Ransomware Threat Hunting and Intelligence

Homayoun et al (2020) advocated employing sequential pattern mining algorithms to discover the best attributes of ransomware programs from benign apps and to identify ransomware software. The efficiency of their detection characteristics was examined by using them with the J48, random forest, bagging, and MLP classification algorithms. The criteria for this investigation were the usual types of True Positive (for total samples now recognized), False Positives (mistakenly identified samples), True Negative (number of correctly rejected samples), False Negative (number of incorrectly rejected samples) (Incorrectly rejected samples).

They begin by identifying and defining detectable patterns and occurrences to identify the appropriate attributes for classification. The sequence pattern mining approach will next be applied to each dataset in order to identify the best sequence pattern. Each sequence in each dataset is then cross-matched based on the maximum sequence pattern to highlight the characteristics of the training classifiers. The following are examples and descriptions of maximum sequence patterns: 1) R (for all events must be registry), 2) D (all events must be DLL), 3) F (all events must be file), 4) RF (multiple transitions, but the first transition is from the registry to the file event), 5) RD (multiple transitions, but the first transition is from the registry to the DLL event), 6) FR (multiple transitions, but the first transition is from file to registry event), 7) FD (more than one transition, although the initial transition is (more than one transition, but the first transition is from DLL to file event).

2.2 Classifications Based on File System/Process Monitoring

Prior attempts to detect malware have primarily focused on monitoring its low-level file system operations. UNVEIL by Kharraz et al (2016) is one such technique. UNVEIL detects ransomware by attempting to monitor file activity. They were divided into three types based on file system operations (whether a file was read, written/encrypted, deleted, or overwritten) to monitor these actions. They may be able to detect ransomware assaults as a result of this. The Redemption by Kharraz et al (2017) method was also used to examine the request pattern of a file I/O to see whether there was any potential ransomware for each process. If this is restored, the processes that have been labelled as dangerous will be terminated. These are good solutions in general, but their drawback is that many harmless apps, such as encryption and compression of applications, also have such file access characteristic features. If this is the case, these solutions risk producing a large number of false-positive findings since they consider those features to be the same when identifying the activity of various ransomware file systems. Below are a few more characterizations based on file system/process monitoring.

2.2.1 RWGUARD: A Real-Time Detection System Against Cryptographic Ransomware

Mehnaz et al (2018) presented RWGUARD a decoy-based ransomware technique rigorously tested to analyze 14 of the most common ransomware and detect their operations in real-time. It used both the file change and the process change to identify files encrypted by ransomware. Three monitoring strategies were used in this approach: file change monitoring, decoy monitoring, and process monitoring. By employing the correct CryptoAPI function and learning characteristics identical to the user's encrypted file, it was possible to distinguish between a benign and an encrypted ransomware file. Using this comprehensive decoy system, it was nearly hard for ransomware to distinguish their fake files.

2.2.2 RANSOMSPECTOR

This method is based on the virtual machine introspection approach which was presented by Garfinkel and Rosenblum (2003). This solution integrates file operations like opening, renaming, closing, reading, and writing with network operations like connecting, binding, receiving, sending, and disconnecting. They then matched them to discover which corresponded to specific system calls in the kernel of the operating system. The virtual machine can also collect this, which includes context information like the system call's return value, the parameters, and, lastly, the caller's process. Tang et al (2020) also discovered that a significant number of crypto-ransomware samples linked to a network create a huge number of network patterns with similar patterns that differ from their file activities. As a result, by analyzing how these ransomware programs interface with the network and file system, they will gain a bit more precision and inform the user if there is evidence of a ransomware assault.

2.2.3 Crypto Ransomware Analysis and Detection Using Process Monitor

Kardile et al (2017) suggested a method for identifying ransomware attacks based on a process monitor implemented on top of Cuckoo Sandbox. They intentionally picked sandbox because it removes the danger of data loss because Cuckoo sandbox returns to its original state after the malicious sample has been executed. This method first builds a genuine and bogus environment to run these ransomware strains. They may then capture the file system calls trail and record the I/O access using a process monitor. Their research discovered that when suspected malware attacked the system being targeted, the behaviors and activities of files in the system altered dramatically. They found that the time stamp for the entries in the Master File table was quite close to a ransomware assault occurring on that system by observing the Master File Table.

2.2.4 Cryptodrop

Scaife et al (2016) presented a solution designed for the Windows operating system, which has been known to be frequently targeted by ransomware. This method of identifying ransomware relies on indicators to track the many ways in which a file is changes. If all of these indicators are discovered to be true in a file, it may be determined that the file has been corrupted and contains harmful elements. These signs include categorizing ransomware activity based on their actions into three categories. For class one, the ransomware would try to rewrite what was in the original file by opening it, reading it, encrypting it, and closing it. For class two, alter the location of the user's file, read and encrypt the file, and then return it to its original position. The file name may change from the original name by shifting the file back and forth. The ransomware would examine the file, produce an encrypted copy, and then destroy or replace the original for the final class.

2.2.5 SSD-assisted Ransomware Detection and Data Recovery Techniques

On the storage side, SSD insider++ method presented by Baek et al (2020) incorporates sophisticated features like online ransomware detection, flawless data recovery, and sluggish detection. The technique described for online detection is one of the key distinctions between this suggested approach and signature-based alternatives. The algorithm watches and analyzes the host machine's I/O pattern and makes a judgment during run time by analyzing invariant traits that characterize the I/O behavior of ransomware-affected host computers. This is especially significant since it now allows for identifying ransomware attacks in their early phases. The SSD-insider++ overcomes the drawbacks of earlier software and hardware in detecting ransomware by combining a ransomware detection and a data recovery algorithm onto a single SSD.

The SSD-architectural insider++'s architecture comprises ransomware detection and backup/recovery. To identify any abnormal behavior, the SSD Insider++ employs two distinct file operations known as "update-after-read" and "trim-after-read." When ransomware attacks files, its goal is to remain undetected for the longest time feasible by the user. As a result, if many I/O patterns are discovered, we can interpret this as a symptom of a ransomware assault. Baek et al. studied the behavior of six prominent real-world malware to capture ransomware behaviors. Zerber, Locky, Cryptoshield, WannaCry, Mole, and Jaff are examples of malware among them. To identify the traits capable of differentiating this malware by comparing their I/O footprints to those of common apps. After training and testing with various combinations of this ransomware and programs, the SSD insider++ was able to identify new or undiscovered malware by recognizing their distinctive I/O patterns.

2.3 Classifications Based on the Network Traffic

2.3.1 The Case of BadRabbit

Alotaibi et al (2021) developed a technique for detecting efforts to distribute ransomware at the network level rather than preventing the device from being encrypted, which was already addressed in prior studies using badRabbit as a case study. To accomplish this analysis, they employ two VMs, one Windows 10 and one REMnux, for static analysis. They operated four virtual machines for the dynamic analysis: one with a REMnux acting as a gateway, two with Windows 10 (one infected with BadRabbit), and one with Windows 7. The investigation showed that Bad Rabbit did not need to contact the command-and-control server to exchange an encryption key; instead, it accessed these files using a public key. Because Bad Rabbit is self-propagating ransomware, our solution employs five modules to identify and fight self-propagating malware. Deep packet inspection (dpi) and packet header inspection are examples of these modules (phi), honey pot-based, ARP scanning-based detection, and SMB Packet size checkers.

2.3.2 Ransomware Early Detection by the Analysis of File Sharing Traffic

Morato et al (2018) presented Ransomware Early Detection from File SHaring traffic which is usually referred to as REDFISH. This solution may be regarded as a framework for detecting and blocking different ransomware behaviors when the infection encrypts data on a network volume from a Network Attached Storage.

This solution examines the difference in traffic behavior between infected and non-infected hosts. The characteristics they investigated for these host behaviors include how files on a shared file are opened, read, written, and deleted. Their approach is derived from studying SMB/SMB2 traffic over a single TCP connection.

3. Proposed Approach

For this work, our major goal is to implement a novel approach to detect the Thano Malware. The analysis would be carried out in three major layers. The first two layers would be the behavioral and signature-based approaches respectively. A figurative illustration is presented in Figure 1. By using both layers, we would leverage the benefits that both malware strategies implement. Layer three acts as a bonus layer to catch any pitfalls from the previous two layers. In this layer, we introduce machine learning techniques to enable us to train the malware classifiers. Figure 2, Diagrammatically depict the procedure of all the layers. We can completely optimize all three approaches for their resistance in identifying malware samples and run-time efficiency by adding the machine learning algorithm layer.

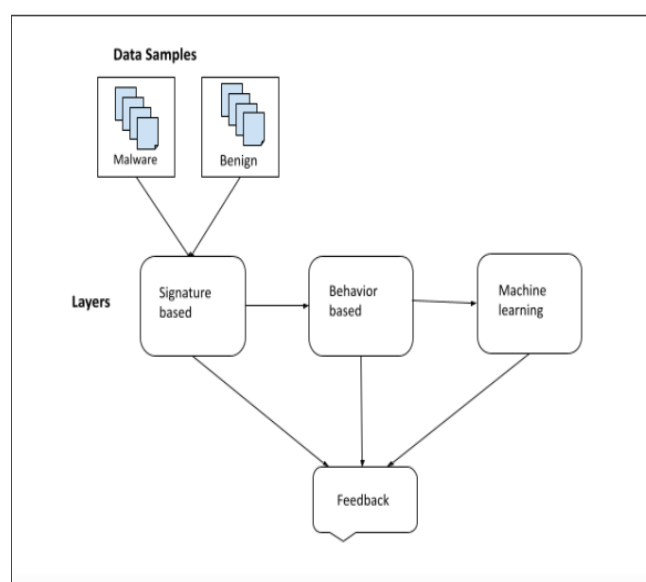


Figure 1: Diagram of the Architectural layer of the hybrid layer system

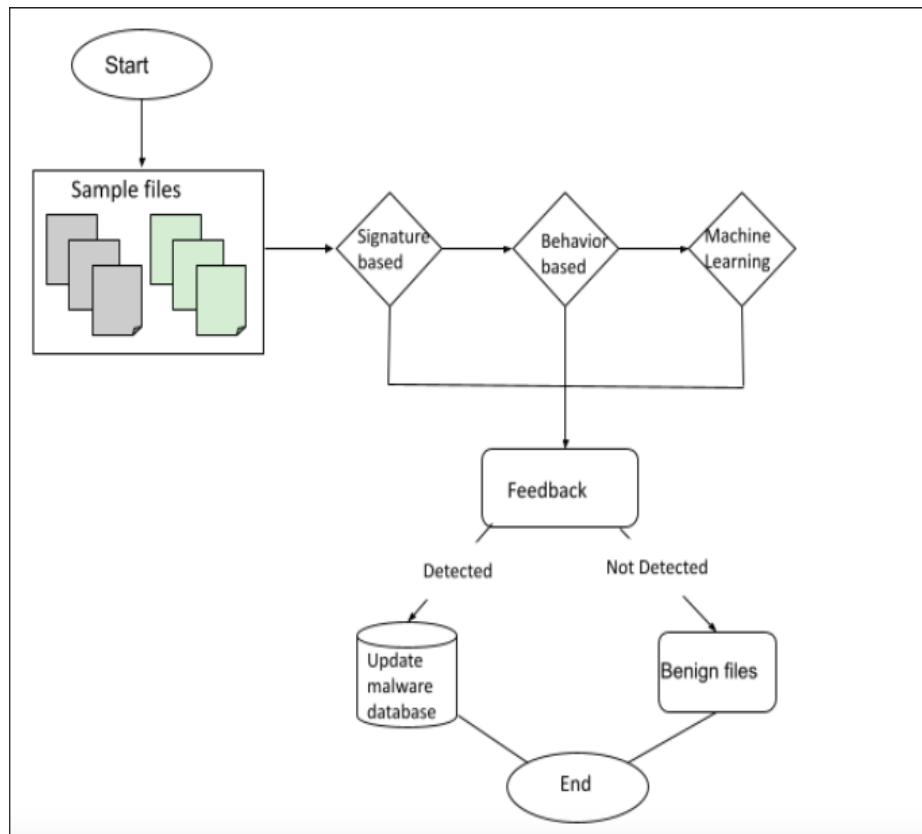


Figure 2: Diagram of the processes of the proposed hybrid technique

4. Research Plan

The primary purpose of this research is to present multiple potential approaches for training malware classifiers using machine learning. Another objective is to prevent ransomware from spreading laterally inside a network. here are two types of malware analysis techniques: signature-based malware analysis and dynamic malware analysis. Each of these strategies has advantages and limitations; the main problem of signature-based detection is that new and undiscovered malware cannot be identified, whereas Dynamic analysis takes time and is rigid. We may utilize the strengths of both analytical approaches while reducing their limitations by combining them to build a hybrid system. To prevent the lateral movement of ransomware, we will use various techniques such as SDN, Moving Target Defense, and Zero Trust.

The objectives are:

- To be able to obtain malware sample data set to perform experiments.
- To be able to properly detect malware, either those that are known or even the new ones.
- To be able to use machine learning algorithms to differentiate between benign and malicious programs.
- To be able to prevent the lateral movement of ransomware.

4.1 Data collection

In previous methods, data were obtained from various websites such as VirusTotal, Virus shares, Zelster, MWanalysis.org, Vxheaven, PCHome Malware Repositories, etc. For research purposes, three major types of datasets are available/ being used and they are:

- Publicly available datasets: These are currently being offered and provided publicly. They are also being updated and maintained by research enthusiasts for the purpose of research all over the world for free in the area of cyber security.
- Artificially generated datasets: These are classified as datasets generated manually using special tools or collected from the network traffic by cyber security researchers.
- Commercial datasets: As the name implies, these are datasets that are not freely offered to the public for use. They are provided as commercial projects and supported by companies for commercial purposes.

For the sake of this research, we shall be obtaining our dataset from publicly available datasets. This is because public datasets are freely available, generate new insights into data collected by fellow researchers, and possess a larger sample size. In order to obtain the biggest sample data, we will obtain malicious ransomware samples from public sites such as virusTotal, Vx heavens, NetLux, Anubis, nexginre for malicious data and Benign samples from portableapps.com. By collecting ransomware sample datasets from a wide variety of publicly available datasets, we can achieve one of our goals of building not only the biggest sample but also the more diverse dataset.

4.2 Data Processing

A testbed similar to what is presented by Kardile et al (2017) will be used. In this testbed, two environments are used. The first is a realistic user environment and the second is an artificial environment. We can then proceed to run the malicious sample in these two different environments. These two environments are built on the Cuckoo sandbox because it is impossible to lose users' data because it reverts to normal after the malicious sample has been completely run. After this, the file system call trace and the I/O access are being recorded using the process monitor. Finally, the proposed system would then detect and analyze the ransomware through manipulating the process monitor. It is also important to note that a virtual machine is often being set up for many dynamic analyses such as this. The need for a virtual machine is to emulate the regular Host OS. Since malicious samples can harm the host OS, a virtual machine OS is being used in its place because even though it looks identical, it is completely isolated from the host OS. Two of the most commonly used virtual machines are VM Workstation and Oracle Virtual Box. Other tools involved in the monitoring of ransomware behaviors include;

- ProcessMonitor: This is a tool that is available on Windows, and it is used for monitoring file systems, processes, threads, networks, and Registry.
- Ollydbg: this is an x86 debugger that is useful for examining the execution of another program.
- WireShark: This is an open-source sniffing and analysis tool that captures network traffic.
- Netcat: Networking tool that can monitor data transfer over a network.

We are combining both the static and dynamic methods of analysis for our hybrid method to provide a more hybrid solution. This hybrid solution is then tested again using four different machine learning algorithms to find the best and most accurate malware detection. The tools we can use for the static analysis in order to be able to extract the static features such as the string, the imports, the exports, and much more, include IDA pro disassembler, capstone, Peid, PsStudios. For the dynamic extraction, we can use tools such as ProcMon, Ollydbg, Wireshark, RegShot, and Cuckoo Sandbox to gather runtime features. As stated above for the dynamic analysis, we would use the OS of a virtual machine so as not to endanger our host OS. Our testbed is being set up this way because while the static analysis has the upper hand by consuming fewer resources and having a higher analysis speed, it cannot detect unknown and obfuscated malware. With the dynamic analysis alone, we can detect unknown malware, but we would consume more resources and find it difficult to detect malware that hides their behaviors during run time. We can catch any fallout from the previous two methods with the additional layer of machine learning algorithms.

4.3 Data Analysis

Classification algorithms are classified into symbolic learning algorithms (CART, C4.5, NewID, AC2, ITrule, Cal5, CN2), statistical algorithms (Naive Bayes, k-nearest neighbor, kernel density, linear discriminant, quadratic discriminant, logistic regression, projection pursuit, Bayesian networks), neural networks (backpropagation, radial basis functions), and Random Forest. However, for the purpose of this research, we will be comparing Logistic Regression, Naive Bayes Classifier, Random Forest, and a Decision Tree.

The aim is to record the performance of Logistic Regression, Naive Bayes, Random Forest, and Decision Tree Classifier on a data set to classify. The result will be tabulated and graphed to show the recommended algorithm for classifying data sets.

The project will consist of the following stages:

- The data will be collected (and cleaned if necessary) from the various open-source ransomware database.
- The data will be separated into training and test data.
- Use the data set to build a decision tree, perform the necessary tests, and record findings.

- Use the data set to train a model using logistic regression, perform the necessary tests, and record findings.
- Use the data set to train the model using Naive Bayes Classifier, perform the necessary tests, and record findings.
- Compare the performance of the algorithms and provide proof (if any) of the recommended algorithm for the input data.

The goal is to train the four selected classification algorithms to predict whether a malware sample is malicious or benign. The data set used for the training and testing will be the same. The performance of all algorithms will be measured, calculated, and compared based on accuracy, speed, and error recorded. The reason for choosing a random forest classifier is that compared to other machine learning classifiers, they are faster and a lot more scalable to deploy. From previous research carried out by Mehnaz et al (2018) for RWGward, it was also noticed that another tree-based classifier known as decision tree had a high performance like the Random Forest for Non-Linear datasets. For these reasons, we can consider using both tree classifiers for non-linear datasets or use the naive Bayes and Logistic regression for others.

4.4 Evaluation

Recent work by Scaife et al (2016) shows that instead of testing the program that was making the changes to the file, their method tested the user's data that was being changed. This means that they are able to create an early alert for ransomware based on how the users' data was changing.

We can test our research by obtaining and modifying some ransomware to develop new ransomware. So, these new ransomware tools will have new signatures and different behavior. These ransomware samples would also be collected from various ransomware families to have a wide range of classifications.

The four algorithms will be evaluated against the data provided through the various publicly available datasets. The data has been collected and correctly labelled, thus serving as good training and test data. The evaluation of the algorithms would be based on their speed (performance), accuracy (in %), and their marginal error (in %).

In order to be precise with the performance of all algorithms, they will be compared with different input sizes and measured accordingly. We would also employ the use of Machine learning tools. There are several machine learning tools available, and they are divided into two: python-based tools and Java-based tools. For the python-based tools, we have tools such as Scikit learn, Keras, and Theano. For the Java-Based tool, we have Weka.

4.5 Lateral Movement Prevention

We will set up another testbed with multiple VMs to capture how ransomware moves through the network. We will implement various techniques to prevent the lateral movement of ransomware and compare the results. These methods include SDN, network segmentation, zero trust architecture, and IP hopping by Moving Target Defense methods. We will compare their success in preventing lateral movement of all ransomware in our dataset. The cost of implementation, the effect on the network performance will also be measured and compared.

4.6 Expected Outcome

We can present a rather comprehensive and detailed solution by combining the various methods and features present in the hybrid-based detection system. Also, we can overcome the deception and confusion that comes with just one type of detection technique to a large extent. This way, it would allow us to understand the program better more systematically, comprehensively, and accurately. Based on the results of our second testbed, we can also classify solutions for preventing the lateral movement of ransomware based on their accuracy and network performance.

References

- Alotaibi, Fahad & Vassilakis, Vassilios. (2021). SDN-Based Detection of Self-Propagating Ransomware: The Case of BadRabbit. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3058897.
- Baek, Sungha & Jung, Youngdon & Mohaisen, Aziz & Lee, Sungjin & Nyang, Daehun. (2020). SSD-assisted Ransomware Detection and Data Recovery Techniques. IEEE Transactions on Computers. PP. 1-1. 10.1109/TC.2020.3011214.
- Garfinkel, T., Rosenblum, M., (2003) "A virtual machine introspection-based architecture for intrusion detection." In: Proceedings of the 10th Network and Distributed System Security Symposium, pp. 191–206.

- Homayoun, S., Dehghantanha A., M. Ahmadzadeh, S. Hashemi and R. Khayami (2020), "Know Abnormal, Find Evil: Frequent Pattern Mining for Ransomware Threat Hunting and Intelligence," in *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 2, 1 April-June pp. 341-351, doi: 10.1109/TETC.2017.2756908.
- Kardile, A.B. (2017), *Crypto Ransomware Analysis and Detection Using Process Monitor*. Ph.D. Thesis, The University of Texas at Arlington, Arlington, TX, USA.
- Khan, Firoz & Ncube, Dr & Ramasamy, Lakshmana & Kadry, Seifedine & Nam, Yunyoung. (2020), "A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning", IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3003785.
- Kharaz, A., Arshad, S., Mulliner, C., Robertson, W., Kirda, E. (2016): Unveil: a large-scale, automated approach to detecting ransomware. In: 25th USENIX Security Symposium (USENIX Security 2016), pp. 757–772. USENIX Association, Austin
- Kharraz, A., Kirda, E., (2017). Redemption: Real-time protection against ransomware at end-hosts. In: Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses, pp. 98–119.
- MahdaviFar, S. and Ghorbani, A.A. (2019), Application of deep learning to cybersecurity: A survey, *Neurocomputing* 347 149–176, <http://dx.doi.org/10.1016/j.neucom.2019.02.056>, <http://www.sciencedirect.com/science/article/pii/S0925231219302954>.
- Mehnaz, Shagufta & Mudgerikar, Anand & Bertino, Elisa. (2018). RWGuard: A Real-Time Detection System Against Cryptographic Ransomware: 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, Proceedings. 10.1007/978-3-030-00470-5_6
- Mirjalili, S., Mirjalili S. M, and Lewis A. (2014), "Grey wolf optimizer," *Adv. Eng. Softw.*, Mar, vol. 69, pp. 46–61.
- Morato, Daniel & Berrueta, Eduardo & Magaña, Eduardo & Izal, Mikel. (2018). Ransomware early detection by the analysis of file-sharing traffic. *Journal of Network and Computer Applications*. 124. 10.1016/j.jnca.2018.09.013.
- Scaife, N., Carter, H., Traynor, P., Butler, K.R.B. (2016): Cryptolock (and drop it): stop-ping ransomware attacks on user data. In: 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), June, pp. 303–312. <https://doi.org/10.1109/ICDCS.2016.4627>.
- Tang, Fei & Ma, Boyang & Li, Jinku & Zhang, Fengwei & Su, Jipeng & Ma, Jianfeng. (2020). RansomSpector: An Introspection-Based Approach to Detect Crypto Ransomware. *Computers & Security*. 97. 101997. 10.1016/j.cose.2020.101997.
- Yang, X.S., and Suash Deb (2009), "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC)*, pp. 210–214.
- Zhang, W., Wang, H., He, H. and Liu, P. (2020) DAMBA: Detecting android malware by ORGB analysis, *IEEE Trans. Reliab.* 69 (1) 55–69, <http://dx.doi.org/10.1109/TR.2019.2924677>.

Analysis of Sexual Abuse of Children Online and CAM Investigations in Finland

Johanna Parviainen and Jyri Rajamäki

Laurea University of Applied Sciences, Espoo, Finland

Johanna.Parviainen@student.laurea.fi

Jyri.Rajamaki@laurea.fi

Abstract: Children who use their free time with phones and computers online, interact with the digital environment on a daily basis, where they will often make social contacts with their friends as well as exchange photos and videos with them. However, children often make contact incautiously with new people online and may risk falling victim to sexual harassment or sexual abuse via sexually charged messages, requests to send nude photos or other Child Abuse Materials (CAM). Most of these cases are unreported crimes. Victims do not always understand that they could be victims. Sexual abuse of children endangers children's psychological, physical and social health, and is against the interests of the child and their human rights. The European Commission's Internal Security Fund Police (ISFP) aims to fund projects on fighting cybercrime and child sexual abuse including digital investigations. This paper provides background information for an innovation project intended to get funding from the ISFP. The case study composes up-to-date pictures of sexual abuse of children online in Finland by applying (1) observations of national police officers who have worked with CAM investigations as a tactical, technical or lead investigator to get unwritten knowledge of the challenge of investigations in the future, and (2) earlier research in Finland. Finnish Child Victim Surveys from past years provide a nationally representative sample of the experiences of children between 12 and 15 years of age where online grooming and sexual abuse is reported. In spring 2021, Save the Children Finland published 11-17-year-old children's experiences of and thoughts on online grooming, and its results show grooming to be a common phenomenon and that a portion of children reported contact being made sexual from its very beginning. This case study combines qualitative and quantitative methods for achieving different aspects and paradigms over the CAM investigations in Finland and are now up to date and would describe how crime investigators have to observe CAM crimes in crime investigations and how cases would be investigated effectively and reliably in a timely fashion and without unnecessary mental workload.

Keywords: child abuse, child pornography, cyberbullying, cybercrime, CAM, Child Abuse Material, case study research

1. Introduction

Child abuse/grooming, child pornography and cyberbullying are the most dangerous cybercrimes (EUROPOL, 2019). Raised awareness can be given to the new exposure that the COVID-19 situation has created with respect to underage people using digital technology. Issues of CAM materials (Child Abuse Material or CSAM; Child Sexual Abuse Material) are meaningful in this phenomenon in general but are also a part of criminal investigations. Other meaningful issues include how children and young people normally share photos and videos even without any sexual meaning behind them. This is not only dependent on Free Internet's possibilities, it is dependent on the Dark Web by anonymous and encrypted sites and new platforms in the future.

2. Research Method and Materials

This study follows Yin's (2014) case study research guidelines and explores what kind of childhood sexual abuse online is seen at open sources. The study considers cases as a crime and as a phenomenon including reported, unreported and hidden crimes as a cybercrime and as a crime against children and their health and welfare. Research questions address *how different forms of contact are made, the nature of each form and the kind of consequences these initiations of contact can have online or in a child's real life.*

The research material of the case study consists of (1) observations of the main author during previous work as a law enforcement authority (LEA) with experience in reported crimes and the related judicial processes, and (2) the previous Finnish studies and reports listed in Table 1.

By getting knowledge widely from different sources, including official authorities whose purpose is to protect children, general purposes in society and legislation, earlier research and different knowledge of voluntary sector reports and surveys by self-informed cases, it is possible to catch an overall holistic picture of child sexual abuse online. After this, it is able to achieve different aspects and paradigms over child sexual abuse online as a crime and recognise the necessity of CAM material investigations in Finland now being up to date. Understanding the phenomenon's complexity can also help authorities to get better results and give ideas to handle and observe cases effectively as criminal processes. This databased point of view can help the LEA to

handle its own mental workload with actual knowledge and recognise better opportunities to find new tools for CAM investigations as necessary digital evidence of investigations but also to prevent and uncover crimes.

Table 1: Research materials

Study	Year(s)
Finnish Criminal Code; Chapter 17, sections 18-22 and Chapter 20, sections 6-7b, 8a-8c and 10-11	1998, 2004, 2006
Finnish Child Victim Surveys from the years	1988, 2008, 2013, 2018
School Health Promotion studies	2017
The reports of the Security Committee	2013, 2017, 2019
Government Report on Internal Security, Publications of the Finnish Government 48, Ministry of the Interior	2021
Rikostorjunnan tila, selvitysraportti 1/2018, Poliisihallitus	2018
Ellonen, Noora & Rantaeskola, Satu (toim.): Lapsiin kohdistuvien väkivalta- ja seksuaalirikosepäilyjen tutkinta, Poliisiammattikorkeakoulun oppikirjat 24	2016
Surveys and Report of <i>Protect Children Finland</i>	
- Tutut tuntemattomat, Raportti lasten ja nuorten nettituttavuuksista sosiaalisessa mediassa	2010
- Internet ja lasten seksuaalinen hyväksikäyttö	2013
- Lasten seksuaalinen hyväksikäyttö ja digitaalinen media	2016
- Grooming internetissä ja lapsen seksuaalinen hyväksikäyttö	2016
- Grooming lasten silmin, selvitys 11-17-vuotiaiden lasten ja nuorten kokemasta groomingista netissä	2021
- Miten suojella lasta verkossa?	2021
- ReDirection project	2021

3. Results

3.1 Child's sexual abuse as a phenomenon and as a crime in Finland

Generally, Cyber-dependent crimes as Cybercrimes either Cyber-enabled crimes are international crimes, which have increasingly larger statistics. The general development of digitalisation or technologies provide for multiple ways to cause harm or crimes. (Jämsén, 2020). Child sexual abuse online is being pictured this same way. The internet has also changed the nature of these crimes. Previously, most of children's sexual abuse was carried out by a person the child knew who was also known in the child's own environment, but on the web, children meet unknown people outside of their own life's real environment - even the digital environment is an important part of children's environment today (Pelastakaa Lapset ry, 2013). National Finnish Child Victim Surveys from earlier years reports that sexual abuse of children has decreased from the 1980s, but from the year 2008 especially, the frequency of girls' first sexual experiences being with unknown persons has increased by one-third overall and in the year 2013 reached almost 40 per cent. In boys, sexual experiences have decreased from 1988 at 10 years from 35 per cent to 21 per cent (Ellonen & Sariola, 2008). The conclusion is that even though sexual abuse has decreased, this has not meant that the risk factors for sexual intercourse among young people have been linearly reduced, but events are more risky because young people often meet strangers. Children and young people have increasing knowledge of sexual abuse and sexual self-determination, which also can show that sexual abuse cases have decreased, but increasing scrutiny of the results of child sexual abuse cases can inform us about better control of authorities and increasing the risk of being caught (Fagerlund, et al., 2013). This has concluded especially during the last 10 years (Sisäministeriö, 2021).

Sexual violence means use of force the wrong way and it addresses violence to individual sexual privilege and self-determination. Sexual violence includes any manner of harmful act or attempt to engage another person's sexuality, unwanted comments or initiation of contact without that person's consent in every manner of relationship. Other's lines can be broken by touching, harassment or force where a situation can include, for example, physical power, psychological intimidation, extortion or threats. Sexual harassment means unwanted suggestions or touching even if the other has shared that another person's behaviour is nonconformist when one important point of this view is also a victim's own experience. On the internet, worldwide, child sexual abuse does not necessarily include physical contact and it can happen by sending sexually explicit messages, photos or videos, children or adults sexually performing by web camera contact, to get child to watch sexually activities or pornography (Pelastakaa Lapset ry, 2013). Additionally, an adult can request sexually explicit photos or videos from a child or add sexual comments to messages or posted photos. (Juusola, et al., 2021)

Some adults are only interested in by CAM material and usually many of them has at the same time many victims or contacts with many children (Pelastakaa Lapset ry, 2013). In a situation where a picture depicting a child has begun to spread across the internet may be equally harmful to children as other acts online or physical acts (Poliisi, 2021). For a child, fear of a sensitive picture or video being shared many times makes this problem more multidimensional generally, but it also increases harm to children or young people, which also impacts on their recovery of what has happened (Pelastakaa Lapset ry, 2013). Most cases of child sexual abuse and child sexual abuse online are unreported crimes just as many cybercrimes are. Victims do not even always understand that they are indeed victims. Additionally, CAM material includes lots of unreported and hidden criminality because of many of those who produce and use child-depicting pictures or photos can act on their own networks on the Dark Web's encrypted sites, which also means that acts can re-occur after long time when users can have many victims (Tanttari, 2018). This autumn, the latest published survey has been released, the approach of which has been to collect information from CSAM users and prevent these crimes on the Dark Web from being reported. CSAM materials report that almost 70 per cent have seen CSAM materials for the first time whilst under 18 years of age and almost 40 per cent have seen such material whilst under the age of 13 (Insoll, et al., 2021).

Today we see increasing child sexual abuse associated with the online universe where both girls and boys are victims on all platforms that children and youths utilise (Poliisi, 2021). Children and youths are victim of some adults' harmful intention and no one platform is seen as protecting a child or youth from having an inappropriate or sexually interaction. In this is way, we also need to understand from that development on how children and young people act online as independent users (Pelastakaa Lapset ry, 2013). Violence and sexual abuse affect children and youth welfare (Fagerlund , et al., 2013), especially their psychic, physical and social health, which is against the interests of the child, the child's rights and can have an influence on familial life (Ellonen & Rantaeskola, 2016).

In the Criminal Code of Finland, 'child' means a person under 18 years old. Sexual abuse of children includes touching, sexual intercourse with child or otherwise performing a sexual act on a child or cajoling a child into an act that can impair his or her development and violate the child's sexual self-determination. (Rikoslaki 20 luku, 2021) For example, possession of a sexually offensive picture depicting a child, the purchase of sexual services from a young person and solicitation of a child for sexual purposes ("grooming") under the age of eighteen are also criminalised (Rikoslaki 17 luku, 2021). Voluntary dating of and sending sexual messages, pictures or videos to those under this same age is not criminalised if there is no great difference in the youth's mental and physical maturity (Rikoslaki 20 luku, 2021). On the other hand, a 15-year-old youth is committing a crime if they send another child or youth sexually explicit photos or post them (Poliisi, 2021). Whether or not that crime is sexual abuse or, for example, mischief as a defamation or dissemination of information violating personal privacy depends on case-based definitions.

The phenomenon is not a new one, online grooming especially is seen as a frequent and increasing phenomenon all over the world. Usually grooming has meant adults making contact making with a child over a period of time where is aim to first build a trusting relationship with the child and after that to aim sexual abuse in either the online or real environment. In conversations the adult may send sexually explicit content to the child, ask the child to send nude photos or ask the child not to tell others about the contact. (Juusola, et al., 2021) Contact making can also happen directly and quickly, including proposing (Kalliokoski, et al., 2021). Children have reported that contact making maybe sexual from the beginning or at least quite quickly after that (Pelastakaa Lapset ry, 2013), (Juusola, et al., 2021). Children can find discussion easy and ingenuous online with strangers because interaction is anonymous and includes physical distance. Conversations can be so intensive that a child does not have normal control over his or her own behaviour even if the child can feel free to halt a discussion if he or she wants to. On the internet, adults can also confirm for children that discussion is normal and confidential and therefore the child can lose control and cannot understand the risk of an adult's behaviour being harmful or criminalised. Children or youths can keep on going with risk-based experience just not thinking about possible consequences in order to feel acceptance or gain understanding. Conversations can go on regarding a child's natural age-driven interest in sexual issues or they may just be curious. It has also been reported that taking part in conversions can be done just by exciting or wasting time. (Pelastakaa Lapset ry, 2013)

3.2 More details of earlier research and surveys

In general, victims of sexual crimes feels guilty about their own victimhood, which can lead to them not reporting crimes. Thus, it has been important to report crimes against children even if it feels minor due to the nature of these crimes. Because of the reporting of cases, it has been possible to recognise victims, to prevent new acts

and stop already ongoing contact making, other acts or crimes. (Poliisi, 2019) Legal Enforced Authority has knowledge of crime statistics and phenomenon by their work experience. It has been recognised when we are talking about unreported or hidden crimes, it is very important to get more specific and better overall information on real phenomena with self-reported methods (Fagerlund , et al., 2013), (Ikonen & Halme, 2018). People have answered quite open questions on their welfare and experience of violence (Parikka, et al., 2018) which make it possible to understand the phenomenon of violence in society and prevents violence in different ways (October, 2018).

Finland has examined Child Victim Surveys from the years 1988, 2008 and 2013. These surveys from past years provide a nationally representative sample of what children between the ages of 12 and 15 experience with regard to online grooming and sexual abuse. The year 2013 survey repeats the questions posed to victims in 1988 and 2008, which makes it possible to compare results to general information on children and young people's sexual experience and sexual abuse by people five years older than they are. Specific questions have been asked of those who have reported experiences. In the year 1988, amongst 14-15-year-old girls, roughly 14 per cent reported sexual experiences with adults, while the results of the same survey in 2008 with girls yielded a frequency of 7 per cent and another such survey in 2013 yielded a frequency of 4 per cent. Boys of the same age first yielded a frequency of sexual experiences with adults of five per cent, while a later survey yielded a frequency of only one per cent. While earlier years yielded more physical sexual experiences than after 2008, the reported experiences that were mostly centred around touching. Most children have not reported feeling that sexual experiences with an adult are abuse. Every year, children or youths have yielded roughly one per cent frequency of an adult interacting with them sexually by creating a sense of obligation, via violence or by giving some form of gifts in exchange for sex. Only a minority has reported that they have met with a previously unknown person but still there has been reported several experiences. (Fagerlund , et al., 2013)

When children's own environment becomes wider with the child's own age and increasing independence the likelihood of that a child can meet violence via a previously unknown person also increases. Part of this threat includes the development of technology, which has brought the internet as a meaningfully contact-making environment for children and youths. Children and youths use the digital environment daily to keep in touch their friends, exchange their contact information and make new friends or other relationships without a face-to-face meeting. In this environment, children and youths feel general negative experiences such as having rumours spread about them, bullying, general spite as well as threatening behaviour (Ellonen, et al., 2008). In the year 2008, about 25 per cent of 15-year-old girls has reported that they have met face to face with a person they have got to know online. Over 20 per cent of girls have shared that they had been asked for sexually explicit photos or received suggestions to have sex. Eight percent of 12-year-old girls have been asked for their photos by an unknown person online. Hiring a person for sex has been reported rarely (Salmi, 2008). Bullying, harassment and asking for sexual photos has been reportedly decreasing since the year 2008, which can tell us that using different platforms with phones and computers is so general for everyone that also knowledge of these phenomena has also increased, and it helps children, youths and their parents to define them better. The Victim Survey in year 2013 added questions of whether a stranger has asked child or a youth to send them nude photos of themselves or perform in front of web cameras. A total of 15 per cent of girls and five per cent of boys have told others about these experiences (Fagerlund , et al., 2013).

Finland has also explored School Health Promotion studies in different years to children and young people. For example, a survey in 2017 has explored sexual harassment and sexual violence committed upon children and youths last year. Sexual harassment has been found to be more general than sexual violence. Sexual harassment means suggestions, harassment, sexually explicit messages, or calling attention to another person's body or clothing. Sexual violence means involuntary touching, applying pressure to have sex or offering or giving money, alcohol or drugs in exchange for sex. Girls have reported sexual harassment being mostly verbal and they have reportedly experienced it more than boys have. Experiences of sexual harassment have reportedly increased as children get older. Most of these experiences have taken place over the phone, online or in some public place. A small percentage of different ages of youths have shared that they have been touched involuntarily or have experienced touch which has triggered fear. These experiences have included both that someone has touched the child or that a child has felt or been forced to touch someone else's intimate area. School Health Promotion studies in the year 2017 have shown that children and youths have experiences of sexual harassment in all age groups, which were included in surveys. These experiences are also quite general with children who are foreigners, who are in some way limited by their acting possibilities or children who have already been taken into custody. (Ikonen & Halme, 2018)

Jusola et al. (2021) surveyed 11-17-year-old children and young people about their experiences and thoughts regarding internet grooming. The survey revealed that this is a common phenomenon, as summarised in Figure 1. Many children share that online contact from an unknown person may have been suggestive of sexual interest right from the beginning. The survey has noticed that children and young people recognise and understand grooming, especially in clear cases. Most children have reported that sexual messages (92%), requests for nude photos of themselves (70%), or receiving them from someone else (89%) can be grooming. Also, when someone wants to make friends (76%), send gifts (70%), ask for contact information (63%), or if someone starts to plan a meeting or ask for confidential conversations (80%) is understood as to be grooming.

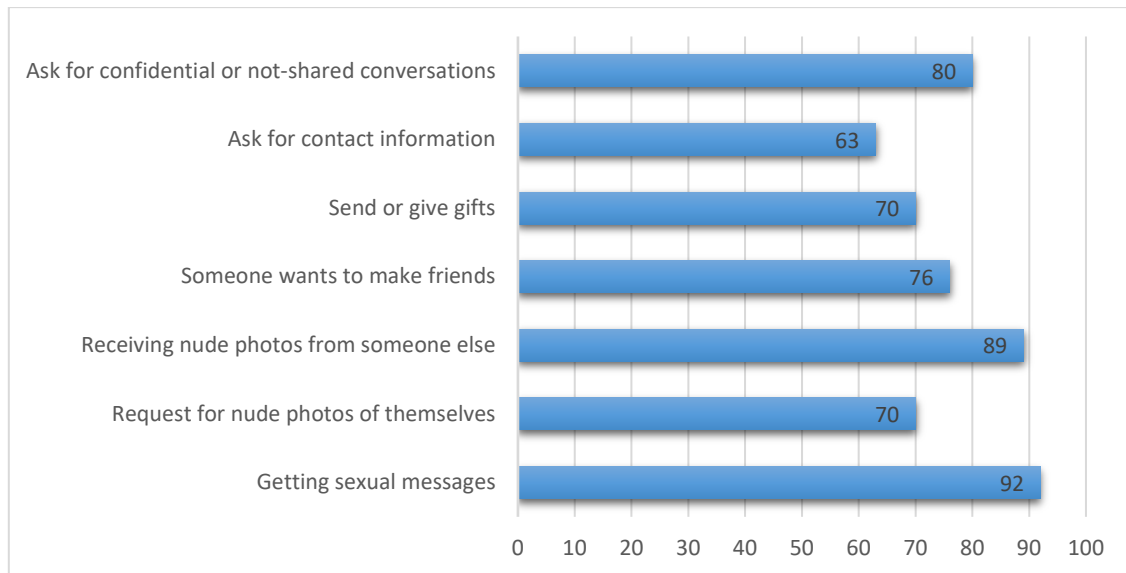


Figure 1: Things that children think are parts of grooming (modified from Juusola, et al., 2021)

Even 73 per cent of children and youths reported that they could recognise a person who is grooming, but at least one quarter cannot say if they could. Children also believe that they can end a conversation which turns sexual but there are also children who have reportedly felt that they would like to end a conversation, but they could not keep in touch or feel attachment to a contact online. Comparing to earlier surveys, 62 per cent of children have reported that they know or they had suspected that the one who initiated contact was an adult or at least five years older. This is a higher frequency than earlier. Only eight per cent of girls and 39 per cent of boys have reported that they have never received sexual messages online, but 17 per cent report that they have received such messages weekly and 29 per cent at least once a month. Almost one-third of girls receive nude photos at least once a month, 15 per cent weekly and boys 12 per cent once in a month and four per cent weekly. Children and youth quite easily and even incautiously initiate conversations with strangers online and provide identification information (name, age, address, hobbies, etc.). Half of answers show that they are used to starting conversation with strangers and it is natural. They can talk about almost everything, such as about daily happenings, shared interesting, hearings, friendship, music, games, sorrows, sex and sexuality. Conversations with unknown are interests (58%), nice (72%), make new friends (71%) or provide support (38%). Contacts are made mainly by dullness (55%) or curiosity (40%) but almost one-third (33%) has not understood what it means when you start a conversation with an unknown person. Six per cent of children have reported that making contact online has led to meeting in the real world. Grooming has caused over half of (54%) feelings of anguish, dejection, depression or other negative feelings from itself but also feelings of embarrassment or guilt. One-third of children have shared that they also felt helpless and one-fifth has felt fear that something awful would happen. If children share about grooming experiences, they usually tell friends or adults but only after a long time. It is meaningful that many wants to tell adults and professionals about their experiences if someone simply asks them. (Juusola, et al., 2021)

3.3 What can authorities do?

It has been especially important to record sexual crimes against children as part of serious crimes against life and health. Victims of these crimes are in a vulnerable position. The judicial process should be more effective and resolved more quickly. That is why an important priority is finding new tools and practices. (Sisäministeriö, 2021) It has already been seen that it is important to protect every child from different violence and for

authorities, it is important to work together that every dimension co-operates to serve the child's benefit (Poliisiammattikorkeakoulu, 2013). There are already many multi-professionals and authorities' working models with official, unofficial or good practices. Police officers, doctors as well as social and health care authorities can consult each other, make contact and change the available knowledge, for example, with schools or kindergarten. Also in pre-trial investigation, it is important to make these contacts and work with prosecutors for better results to guarantee child victims' positions, suspects' rights and an effective judicial process in court. These practices are already normalised because there is needed co-operation and shared professionalism for multiple phenomenon's investigations to get better pictures of the whole situation, guarantee good results for children's ongoing health and social development. That which benefits children is tied to every professional and authority and it is not just a cost-effective act; it is best for everyone and the possibility to recognise and intervene in the same kinds of things in the future. (Ellonen & Rantaeskola, 2016)

Nowadays, criminal investigations quite routinely mean jobs with computers, phones and the internet (Sisäministeriö, 2021) and that is why the challenges of ICT investigations (information and communication technology) have also increased. In 2017, a Laurea master's student wrote his master's thesis on issues of CAM investigations. He interviewed police officers on tactical, technical and leadership investigations to gain unwritten knowledge on the challenges of CAM investigations. These results have shown that CAM investigations are demanding, requiring a lot of time and human resources. Workload, especially mental workload, can be hard or even unreasonable, easily causing cumulative stress for investigators. Investigations based just on using human resources has felt unpleasant and a very hard part of crime investigation. Many investigators have also felt that work that is needed in investigations at the pre-trial phase has had no value or positions in court or at least punishments have been quite small compared to this workload. It has been important to get better tools and opportunities to answer technical challenges and development, for example, with automatised photo and video material classification. There has especially been expectations for a hash-database, the knowledge of which could help a lot in later investigations in terms of making them easier, quicker and erasing extra exposure for harmful materials because of the same circumflex pictures, photos or videos should already be recognised and the knowledge could be used again. In 2018, the National Bureau of Investigations has shared that this kind of database is an ongoing development process and they have co-operated with operators to ensure that sites that contain illegal materials could be closed easily (Tanttari, 2018). Tactical investigators have also earlier reported, with regard to the needs of education on information technology, forensic IT tools and knowledge of different platforms or other technical issues. Many investigators get the best knowledge just by work experience over time. Investigators have also reported that there is a need to gain knowledge which addresses what has happened with illegal materials, for example, how they have come into a suspect's possession and what components should be taken into consideration in interrogations (Marttila, 2017). The same viewpoint has also repeated a survey of the Finnish institute of health and welfare in which has explored authorities' co-operations and practices in 2019 (Huittinen, 2019).

3.4 Old and new challenges to be aware of child's sexual abuse online

It is also considered that cybercrimes are increasing by using new forms of networks - and acting together in different hardly encrypted anonymous marketplaces on the Dark Web. At the same time, it has been recognised that authorities can impact and respond to a cyber environment's challenges only by using actual development and act both nationally and internationally. As earlier, effectiveness is also important such that police could prevent, recover and make crimes' investigations during pre-trial quickly (Turvallisuuskomitea, 2017a). Legal Enforcement Authorities' jurisdiction and resources have to respond to changes of environment and duties' demands also motivated human resources has seen to be adding national cybersecurity and cost-effective actors (Turvallisuuskomitea, 2017). When we talk about child sexual abuse, it is also an issue of border crossing crimes and investigations which require international police and judicial co-operation to get evidence which is needed for investigations. National legislation should be clear and able to respond different challenges. Handling and analysing digital evidence are also important challenges for the cybercrimes area (Turvallisuuskomitea, 2017a). New technology can also free human resources to get better results in IT crimes investigations and they could be cost effective in taking care of possible distractions with worthwhile, authorised quality demands (Sisäministeriö, 2021). In crimes entailing the sexual abuse of children online, there should be tools to catch those who have already committed acts online a long time and also are effective in this way even if there are also a lot of cases where those who are suspects at first time already have a lot of victims (Tanttari, 2018). It is not only an issue of IT forensics, but also that apart from pre-trial and crime processes, authorities could also have enough knowledge of cybersecurity technology, knowledge technology and information technology at all levels of the judicial processes (Ulkuniemi, et al., 2018). It is also important to be aware of what happens as

technology changes and how children and young people use different technological opportunities and what changes or places are relevant now and in the future to this point of view and that various cases of sexual abuse of a child could be prevented and investigated in the best possible manner (Fagerlund, et al., 2013). For example, only nine per cent of children reported in spring 2021 that they use video contact at all or were asked to perform in front of a web camera (Juusola, et al., 2021) even if previous years have estimated that with CAM materials, there is also the increasing issue of live streaming performances so that victims and suspects can be on the other side of the world and there is no recording of that performance whatsoever (Tanttari, 2018). All this could be authorities being able to be prepared for and predict, which is also an issue of necessary ICT investigations' forensic tools in the future to handle different media, new equipment and frequently updated versions (Marttila, 2017) and also be prepared for other issues, which can assist in delivering better results overall.

4. Discussion

This case study pictures what kind of phenomenon the sexual abuse of children online is, what it includes, and how it could be handled better. The analysis of Finnish LEAs' and voluntary sectors' studies and reports provide quite wide components concerning how sexual abuse of children online can get started and by how many different avenues it can happen. A broad point of view is important for understanding the phenomenon, which includes unreported and hidden criminality. Consequences for children's and youth's lives can be, in the worst cases, even devastating if they could have better knowledge to defend themselves in a new living environment with internet access. There are both old and new challenges to the LEA, including tools and possibilities to aim to gain better and quicker judicial processes to guarantee every stakeholder's rights and benefits. This entails up-to-date technical tools, judicial knowledge of national and even international legislation, and an understanding of different phenomena which occur in new environments and needs ongoing monitoring of forthcoming threats and possibilities.

More research is needed that compares Finnish knowledge with an international perspective. For example, reference data is available from Australia. The Australian Government annually reports online child sexual exploitation focusing on community awareness levels, current attitudes towards online child sexual exploitation, and preventative measures. The Australian Centre to Counter Child Exploitation (ACCCE) has interviewed parents or caregivers of children and youth ages 4 to 18, their educators, siblings, and other key players such as health care professionals and other community members. According to ACCCE (2021), 65 per cent of parents and caregivers perceive their children are safe using the internet. On the other hand, 21 per cent of parents and caregivers believe online child sexual exploitation could also happen. The Australian Government's Online Safety Act 2021 improves statutory schemes, functions, and powers requiring, for example, industries to keep their users safe online by updated industry codes or standards. The Cyber Report Team investigates and helps to remove illegal and harmful content from the internet. The eSafety Commissioner can identify anonymous online accounts used to bully, abuse, or humiliate others, or exchange illegal content.

Comparing the experiences of different countries can provide new possibilities and tools for preventing, detecting, or investigating these crimes. At the same time, the use of digital evidence may evolve. Other observations also emerge about how law enforcement technologies have evolved and what will be needed in the future.

References

- ACCCE (2021) Online Child Sexual Exploitation: Understanding community awareness, perceptions, Attitudes and Preventative behaviours, The Australian Centre to Counter Child Exploitation.
- Australian Government (2020) Annual Progress Report: Implementation of recommendations from the Final Report of the Royal Commission into Institutional Responses to Child Sexual Abuse, Commonwealth of Australia.
- Ellonen, N., Kääriäinen, J., Salmi, V. ja Sariola, H. (2008) Lasten ja nuorten väkivaltakokemukset, Tutkimus peruskoulun 6. ja 9. luokan oppilaiden kokemasta väkivallasta, Poliisiammattikorkeakoulu 71/2008, Tampere, Oikeuspoliittinen tutkimuskeskus 87.
- Ellonen, N. ja Rantaeskolä, S. (2016) Lapsiin kohdistuvien väkivalta- ja seksuaalirikosepäilyjen tutkinta, Poliisiammattikorkeakoulu, Tampere.
- Ellonen, N. ja Sariola, H. (2008) Seksuaalinen hyväksikäyttö, Poliisiammattikorkeakoulun raportteja 71/2008, Tampere, Oikeuspoliittisen tutkimuslaitoksen tutkimustiedoksiantoja 87.
- EUROPOL. (2019) Internet Organised Crime Threat Assessment (IOCTA) 2019. European Union Agency for Law Enforcement Cooperation.
- Fagerlund, M. ym. (2013) Lasten ja nuorten väkivaltakokemukset, Poliisiammattikorkeakoulu, Tampere.

- Huittinen, M. (2019) Lapsiin ja nuoriin kohdistuvien väkivaltaepäilyjen tutkiminen lapsen edun mukaisesti, Selvitys poliisin ja perhekeskustoiminnan näkökulmasta, Terveyden ja hyvinvoinnin laitos, Helsinki.
- Ikonen, R. ja Halme, N. (2018) Lasten ja nuorten kokemus seksuaalinen häirintä ja väkivalta, Terveyden ja hyvinvoinnin laitos, Helsinki.
- Insoll, T., Ovaska, A. ja Valkonen-Vaaranen, N. (2021) CSAM Users In The Dark Web, Suojellaan Lapsia ry.
- Juusola, A. ym. (2021) Grooming lasten silmin - selvitys 11-17-vuotiaiden lasten ja nuorten kokemasta groomingista netissä, Pelastakaa Lapset ry.
- Jämsén, C. (2020) Tietoverkkorikollisuus poliisin silmin 2019-2020, Keskusrikospoliisi, Poliisin kybertorjuntakeskus.
- Kalliokoski, E. ym. (2021) Miten suojella lasta verkossa? Opas aikuiselle., Pelastakaa Lapset ry.
- Kurenmaa, T. (2018) Rikostorjunnan selvitystilahanke, loppuraportti 1/2018, Poliisihallitus.
- Marttila, J. (2017) IT-forensiikan tutkintaoppaan kehittäminen CAM-rikostutkiminnan avuksi, Laurea Ammattikorkeakoulu, Kerava.
- October, M. (2018) Lasten, nuorten ja aikuisten kokemuksia väkivallasta, Terveyden ja hyvinvoinnin laitos, Helsinki.
- Parikka, S., Pentala-Nikulainen, O. ja Lounamaa, A. (2018) Kysely- ja väestötutkimuksen tuloksia, Työpaperi 28/2018, Terveyden ja hyvinvoinnin laitos, Helsinki.
- Pelastakaa Lapset ry. (2013) Internet ja lasten seksuaalinen hyväksikäyttö, Ota puheeksi, Pelastakaa Lapset ry.
- Poliisi. (2019) Poliisin ennalta estävän työn strategia 2019-2023, Poliisi.
- Poliisi (2021) Seksuaalirikokset, Poliisi.
- Poliisiammattikorkeakoulu. (2013) Rikostutkinta lapsiin kohdistuvissa väkivalta- ja seksuaalirikoksissa, Poliisiammattikorkeakoulun oppikirjat 20, Tampere.
- Rikoslaki 39/1889. (2021) Rikoslaki 17 luku.
- Rikoslaki 39/1889. (2021) Rikoslaki 20 luku.
- Salmi, V. (2008) Väkivalta sähköisten viestimien kautta, Poliisiammattikorkeakoulu, Tampere.
- Sisäministeriö. (2021) Valtioneuvoston selonteko sisäisestä turvallisuudesta, Sisäministeriö, Helsinki.
- Tanttari, S. (2018) Lasten seksuaalinen hyväksikäyttö saa netissä uusia muotoja, Oikeusministeriö.
- Turvallisuuskomitea. (2017a) Yhteiskunnan turvallisuusstrategia.
- Turvallisuuskomitea. (2017) Suomen kyberturvallisuusstrategian toimeenpano-ohjelma 2017-2020, Turvallisuuskomitea.
- Ulkuniemi, K., Piironen, T. ja Toiviainen, T. (2018) Kyberrikollisuus poliisin haasteena, Poliisiammattikorkeakoulu, Tampere.
- Yin, R. (2014) Case Study Research Design and Methods, Thousand Oaks, Sage.

Digital Risk Management: Investigating Human-Factor Security with a Behaviorist Approach

Ruan Pretorius and Dewald Blaauw

Stellenbosch University, Cape Town, South Africa

ruanpretorius.rp@gmail.com

dnblaauw@sun.ac.za

Abstract: The successful digitization of modern organizations relies on the cohesion between information technology and the workforce responsible for managing and operating it. Without proper management and operation, even the most sophisticated technologies may become vulnerable when operated by an incompetent worker. Numerous studies acknowledge human vulnerability in cyber security, known as human-factor security, as the “weakest link” in a digitized organization’s security posture (Ani et al., 2019). It was found in literature that there exists a shortage of focus on the impact of human-factor security on information and data security in organizations. Through risk management frameworks, the focus is placed much more on the risks posed by technologies, as on the risks presented by humans implementing, managing or interacting with these technologies. Thus, the need to investigate how risk management frameworks could be applied to human-factor security in digitized organizations arise. This paper offers an in-depth understanding of the behavioral and cognitive science of people in relation to digital threat awareness and response. Thereby, insight is gained on how to decrease the human worker’s vulnerability and susceptibility to cyber threats when interacting with technology onsite and offsite the organization. With the Behaviorist Learning Theory, the security posture of an authentic dataset consisting of South African workers in the digital environment was investigated, enabling the observation of stimuli-response behaviors. Appropriate risk management methodologies were applied to identify, classify, assess, and respond to the risks found in the investigated behavior from the dataset. This research project provides security and risk managers with insight into human vulnerabilities and behavior when interacting with technology. This insight emphasizes the importance of human-factor security in an organization’s security posture. Additionally, this insight enables the enhancement of an organization’s security posture with the emphasis on human-factor security within risk management plans. Through a survey investigation the majority of digital workers, in the exposure to digital threats, show high levels of readiness and awareness by appropriate actions. However, the study and analysis confirm that there is indeed evidence of digital workers who portray risky behavior which can result in devastating consequences regardless of the low probability of occurrence.

Keywords: Digitization, human vulnerability, risk management, information and data security, digital risks, cyber threats

1. Introduction

Cyberthreats aim to exploit the shortcomings within digitized companies. In many cases, the workers can be regarded as one of the weakest links in terms of cybersecurity (Ani et al., 2019). While the advancement of technology allows for more complex and sophisticated systems, various challenges and cyberthreats are introduced. Thus, more people are susceptible to making unintentional mistakes or errors that put the organization at risk. It is far easier to exploit humans than it is to exploit secure information systems or technology. Cyberthreats are magnified when managers of digitized companies overestimate the security of technology, while also neglecting the effects of human error. It is clear that the impact and management of humans are a crucial element in cyber security.

Employees can pose as one of the greatest internal risks to digitized companies. Therefore, the success of digitized organizations and the efficacy of information systems rely greatly on the employees who interact with these systems. The human role in digital risk management is of cardinal importance for organizations to establish a strong security posture.

This research project aims to provide insight into human vulnerabilities and behavior when interacting with information systems and technology in digitized organizations in order to improve risk management. To investigate this vulnerability, data will be collected through surveys provided to employees from digitized companies on the subject of human behavior in information and data security. The Behaviorist Learning Theory will be used in data collection to investigate stimulus-response behaviors when employees interact with information systems or technology. Additionally, risk identification, assessment and management will be applied to the stimuli-response scenarios presented in the survey. Specifically, various threats and vulnerabilities will be identified whereafter risk assessment will be conducted on the various identified threats.

2. Problem statement

The impact and risk that human workers hold in terms of information and data security for an organization is greatly underestimated (Evans et al., n.d.; Nobles, 2018). The focus is placed much more on the risks posed by technologies rather than on the workers implementing, managing or interacting with these technologies. Even the most sophisticated technology can become vulnerable when operated by an incompetent worker. The existing papers focus on the general principles of risk management and how to establish a foundation for effective and successful risk management strategies. None of these papers focus extensively on how risk management is applied to risks presented by human vulnerabilities within digitized companies. Additionally, these frameworks are rarely tested with an authentic dataset. More specifically, none of these studies implement a stimuli-response strategy to determine how people behave when exposed to certain stimuli within their day-to-day, onsite or offsite activities.

An in-depth understanding of the behavioral and cognitive science of people in relation to digital threat awareness and response is needed to gain insight on how to decrease their vulnerability and susceptibility to cyber threats. An in-depth analysis and research of organizational as well as digital risk management strategies and frameworks is needed in order to understand its level of success and effectiveness. From the problem statement, the following research questions arise:

1. Are modern digitized organizations aware of the impact and vulnerability of the workforce related to cyber threats? (RQ1)
2. Do digitized organizations implement risk management strategies that cater for or mitigate potential threats caused by the workforce? (RQ2)
3. Does the average worker possess the level of knowledge, skill and risk awareness to implement adequate security measures in their day-to-day activities within organizations? (RQ3)
4. Is environmental stimuli-behavior investigation effective enough at gaining an understanding of behavior when potentially exposed to cyber/digital threats? (RQ4)

3. Methodology

3.1 Research design

The aim of the research design was to determine the degree of security and digital risk awareness visible among employees working in digitized companies as well as the level of risk and security awareness and readiness from the organization itself. A semi-structured survey was deployed which targeted workers within the modern day digitized (technology driven) organization. Additionally, utilization the Behavioral Learning Theory, the behavior of digital workers when potentially exposed to cyber threats in different environments could be investigated. From this investigation, the various kinds of risks associated with the digital workers' behavior was identified and assessed.

3.2 Survey structure and design

The survey deployed which captured empirical data for a qualitative analysis, was structured into three main sections. The first section (Section A) consisted of multiple-choice questions, mostly in the form of Likert-scales. Section A was structured with questions that relate to four main themes. These themes are formulated as it encompasses a wide variety of factors and influences that are applicable to the modern-day employee where technology is prominent in every aspect of one's life. The first theme asks questions that relate to the participants trust within their organization with regards to the safeguarding of personal information and data. Additionally, this theme includes questions relating to participants experience of the usability of organization information and work device usage rules. The second theme related to password management. This includes password change frequency, reuse and complexity. The third theme related to the protection of personal devices. This included anti-virus usage, update frequency and data backup habits. The fourth theme related to overall technology proficiency and awareness. This included questions that related to participants' normal day to day activities with the usage of technology. Thereby insight could be gained on how aware and cautious participants are of the threats that revolve around everyday usage of technological devices.

The second section (Section B) consisted of open-ended questions with the focus on how the participants respond when exposed to certain stimuli. With this section utilizing the Behavioral Learning Theory, whereby participants provide their personal responses, it imitates an environment where stimuli-response behaviors can be observed. The questions were structured around exposure to stimuli that occur both within the organization

and offsite from the organization. The questions range from how a participant would respond when a virus warning notification appears on their workstation computer, how they would react when experiencing slow internet connection or when they are in need of an internet connection but are at an offsite location. The various responses or reactions to such stimuli exposure could potentially introduce many threats to the organization. For e.g., connecting to an organization's portal with an unsecure connection pose as a common threat as assumably many mobile devices use unsafe connections. When a respondent connects to their company portal with an unsecure connection it creates an exploitable vulnerability.

Similar to Section B, the last section (Section C) also consisted of open-ended questions. The focus of Section C was to investigate personal biases and perceptions relating to participants own posture and abilities in cyber security. Additionally, this section investigated participant's view on the security culture within their organization and what practices they employ to strengthen the over security posture.

3.3 Qualitative study

The goal of the qualitative study was to gain insight and an understanding of the types of threats and risks that arise from the behavior and activities of people who rely on technology in their everyday life. This includes work-related and non-work-related activities. Additionally, this study investigated people's understanding of digital risks and security behavior as well as how they perceive the security culture within their own organization. A QAC-DAS software, "Taguette" was chosen as the tool to conduct the qualitative analysis with, as this software allowed for the grouping and analysis of themes within a set of qualitative data.

The Behavioral Learning Theory was utilized to aid the qualitative study by setting up questions (Survey Section B) that emulate a stimuli-response environment. Thereby, the survey participants' responses when exposed to environmental stimuli, in this case digital threats, could be analyzed. Thereby, insight could be gained into the triggers and causes of certain behaviors. The Behavioral Learning Theory was deemed an appropriate method to be utilized in the survey as this theory states that all behavior can be learned and observed through conditional exposure in a certain environment (Western Governors University, 2020). Survey section B was thus developed, mimicking an investigation of participants' behavior when exposed to threats in certain conditions.

3.4 Participants and procedure

For the target audience of the survey and the qualitative study, any person who forms part of the working class in South Africa who works in the IT industry or related field are applicable. This includes any person who considers themselves heavily dependent on technology in order to do their day-to-day working activities. In terms of participant demographics, there is no specific age requirement, but the participant must be part of the labor market. This assumes that all participants will be older than the age of 18. In total, 58 responses were gathered from the survey. Participants were still prompted to provide their age in order to identify whether there exist various levels of threat and risk awareness, perception and behavior among different age groups. Participants' gender was prompted in order to identify and quantify any potential differences in awareness and behavior among gender groups. The age and gender distribution can be seen in Figure 1.

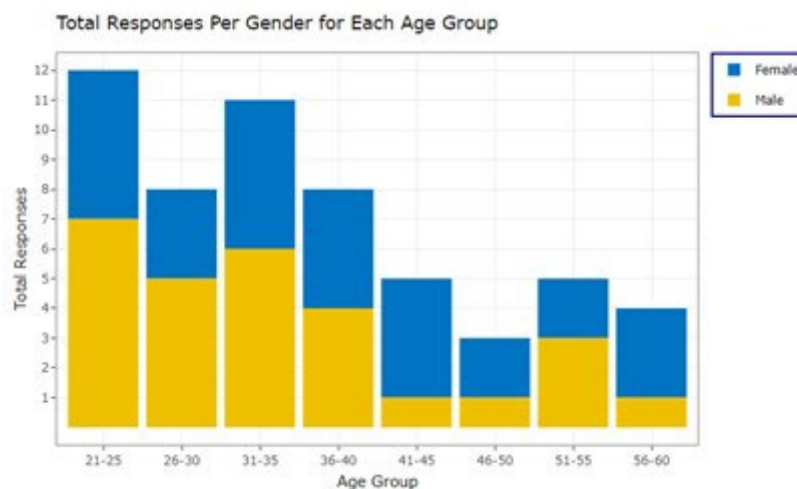


Figure 1: Total responses per age group per gender

Lastly, participants were prompted to specify their current job/role in the labor market. Through the specification of different job roles and titles, there was aimed to determine the general level of threat awareness and ability between different IT or related industries. These general demographic variables serve to identify trends in behavior and patterns among diverse groups in the dataset. It was expected that there is no significant difference among gender groupings. It was assumed that no specific variation in the behavior of gender groups will exist.

3.5 Data collection

Before any participation in the survey is made possible, participants were prompted to provide acknowledgement of the purpose and goal of the survey as well as indicate their consent to participate. For those who do not indicate acknowledgement and consent, the electronic survey was physically unable to start with the questions, ensuring that ethical requirements are met.

The survey was designed, created and managed directly from Google Forms. All data collected from the surveys was stored and secured on Google Forms and exported for analysis. A target with a minimum of 50 respondents was set in order to ensure diversity and accuracy in results as well as increasing the scope of unique answers. The survey was distributed through word of mouth as well as through an invitation message on LinkedIn.

4. Results and analysis

The purpose of the survey study was to investigate the overall level of security awareness and response to threats that are present and that any worker may encounter within the digitized environment. Through the deployment of the survey, it allowed to the capture of responses through the investigation on an actual dataset. Additional emphasis was placed on the actual observable behavior from employees when exposed to certain stimuli. In this case, the stimuli refer to the exposure to any form of security threat. The information found from the investigation aided in gaining a perception of the overall security posture that are already present within the digital workforce. Additionally, the investigation aids in gaining a perception what the type of potential risky behavior are and how to prevent or remediate the threat that may arise from such behavior. From the results of the analysis, the research questions could thus be answered.

As previously discussed, the survey study consisted of three separate sections, all which contribute to the overall understanding of the security posture of the sample. With regards to the gender distribution across the various questions for section A, there is no specific trend indicating a significant variation between Male and Female respondents. This confirms the pre-assumption that there would be no significant variation between gender groups. Within the analysis of survey Section A, it was found that the majority of respondents answers show that there indeed exist security and risk awareness through their interaction and usage of technological devices. This is visible from answers to questions across all four subsections. As can be seen in figure 2 and 3, the majority of participants do back-up their data at some time and the majority are confident that their interaction with technology is in a safe manner.



Figure 2: Data backup regularity per gender group

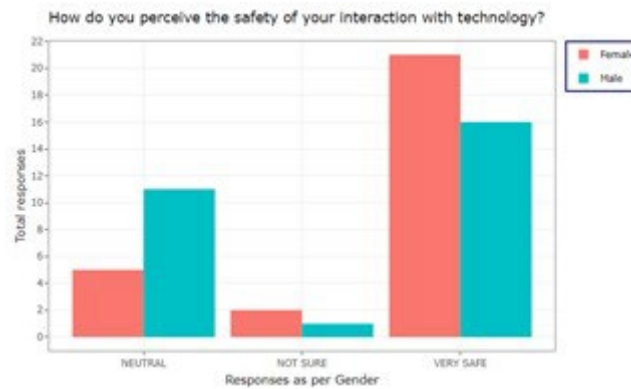


Figure 3: Gender group perception of technology safety usage

However, it is evident that many respondents (the minority) show actions that tend to be less secure or favorable such as with regards to password management. It was found that 11 respondents (19,6%) never update their account passwords. Additionally, it was found that 9 respondents (16,1%) create accounts with easy rememberable passwords, while 31 (56,4%) use repeatable passwords for various accounts as seen in figure 4.

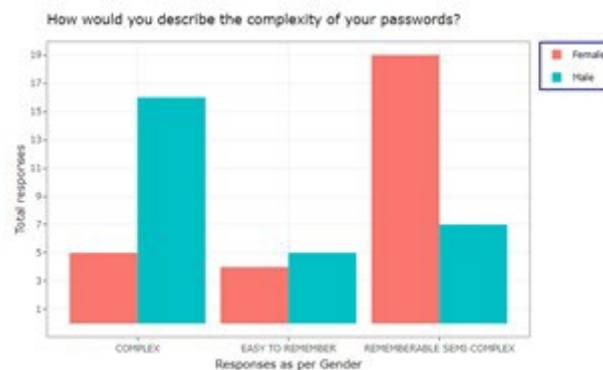


Figure 4: Password complexity per gender group

With regards to device and data security, just below half of respondents do not use anti-virus software on all their devices and four respondents never back-up any of their data as seen in figure 5.

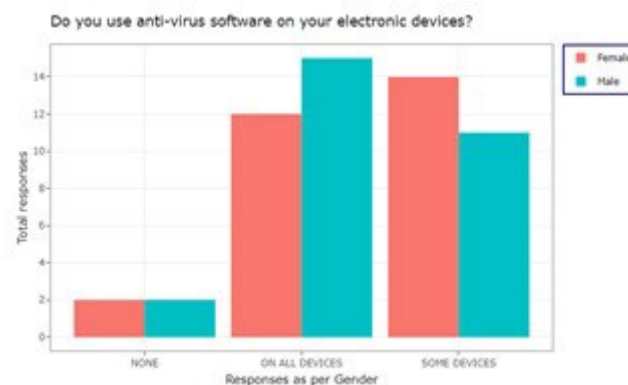


Figure 5: Anti-virus usage

Additionally, 14 Respondents either do not know how to take security precautions or do not feel the need to take precautions when working with technology. 24 Respondents do make use of their work-related Email accounts for personal matters. Additionally, there exist a high admittance to accidentally sending an Email to the wrong recipient (30 respondents). Related results are found for the usage of work-related devices for personal purposes. A high unattendance rate of work-related devices in the workplace was found (34 respondents). This number of respondents contradicts the amount (14) who do not take precautions when working with technology, as leaving a device unattended could be regarded as unsafe. This behavior is influenced

by the level of trust an employee has in the safety of the organization's working environment. Lastly, with regards to EFT purchases, over half of respondents make EFT purchases directly from a third-party website, which was found to be extremely risky (Business Insider SA, 2020). Although for some questions the amount of risky behavior from answers seems to be high, across the entire section, this in general is the minority.

The aim of section survey section B was to identify risks that arise from respondents' behavior when exposed to certain potential threats, which represents exposure to stimuli. Both the response to the exposure to stimuli within the organization as well as offsite from the organization was investigated to broaden the scope of exposure to threats. Within the analysis of this survey section, it is immediately evident that most respondents already portray a strong sense and awareness of security threats and how to react accordingly. This is visible in both subsections of the survey. With regards to stimuli exposure within the organization (Survey Section B1), it appears that when respondents are situated physically within the organization, the evidence of risky behavior decreases significantly. This may be because of respondents being more cautious and aware when they are at work from potentially being under surveillance. Additionally, although not investigated, a physical organizational environment may stimulate focus that could improve caution and awareness as opposed to being offsite where the influence of distraction is present. The higher secure behavior could also be because of infrastructure provided by the organization for onsite use, such as internet connectivity and electronic devices. It was not evident in any respondents' answer that they were allowed to make use of their own data connection for onsite work-related purposes. However, when respondents are situated at an offsite location (Section B2) and in need of a data connection, thereby creating the need for self-provided data, risky behavior starts to occur. It was found that many respondents will make use of self-provided data or public Wi-Fi networks to carry out work related activities, without the mentioning or consideration of the risks associated with such actions. Most respondents will acknowledge the importance and need of secure connections and thus in many cases state the usage of VPNs to secure the connection. An interesting finding only noticed in single respondents was the belief that a certain threat has never or will not occur to them and as such they do not provide a response or answer.

As most respondents show an intense sense of awareness and caution, the identified risks only pose a threat by how easily an employee can be deceived/defrauded for their access authority. The more an employee has access to a certain level of data or information system, the higher the loss or damage can occur from the infiltration or exposure of the data or information system. Thereby, access and authority are identified as a particularly crucial factor for the security of information and data.

In order to assess and classify the risks that were identified in Survey Section B, a risk assessment formula was derived from many existing studies who utilize a standard formula within their risk assessment investigation (Kure et al., 2018; GOVERNANCE & STANDARDS DIVISION, 2017; NRECA, 2011; Ao et al., 2008). The risk formula derived and utilized are as follows:

$$\text{RISK} = \text{LIKELIHOOD} * \text{IMPACT}$$

The likelihood of a threat was determined by three underlying factors: the capability of the threat, the nature of the threat and the effectiveness of current controls in place to prevent an attack. Impact severity of a threat was classified on a scale of one to five, ranging from low to high, as described by GOVERNANCE & STANDARDS DIVISION (2017). The potential worst outcome from a breach that results from all identified risks can lead to very high fiscal damages and repercussions for any organization. As a result, an impact score of 5 was assigned to all identified threats from the results of the survey. After applying the risk assessment formulas to all scenarios presented in Survey Section B, all the identified risks had a probability of occurrence percentage lower than 15%. In combination with the impact assessment of the worst possible outcome for the organization, that is the infiltration of malware, all risks can be classified with a total score of 5. The number is calculated from the multiplication of the impact score of 5 with a probability score of one. Utilizing GOVERNANCE & STANDARDS DIVISION (2017)'s risk assessment matrix (Figure 6), all risks could be regarded as extremely low. However, due to the high impact and consequences of a breach, all potential risks show be avoided where possible.

		Impact				
		1	2	3	4	5
Likelihood	1	1	2	3	4	5
	2	2	4	6	8	10
	3	3	6	9	12	15
	4	4	8	12	16	20
	5	5	10	15	20	25

Figure 6: Risk Impact Matrix (GOVERNANCE & STANDARDS DIVISION, 2017)

Due to the significant fiscal impact that malware can have on organizations, with R3 871 010.33 (+- \$ 261673.57) being the total recovery cost for South African organizations respectively (Sophos, 2020), even the slightest human error should be prevented. It is thus clear that organizations should consider every potential action that an employee may take that could cause a security breach, regardless of the low probability of occurrence.

It was found through an investigation and research in a study that South African digital workers portray a high security posture (Mcanyana et al., 2020). Additionally, this research suggests that South Africa is in the top 5 global countries at preventing data encryption from malware (Mcanyana et al., 2020). In contrast to the beforementioned research, Sophos (2020) proves South Africa's defense capabilities otherwise. Sophos (2020) states that cybersecurity in South Africa is not on par or as robust as other global countries. Additionally, Sophos (2020) continues that as a result of South Africa's substandard security posture, various threat actors consider South Africa as a malware testing ground. These actors test tools and techniques on South Africa, before attempting the deployment on more sophisticated countries (Sophos, 2020).

The aim of survey Section C was to investigate respondents' personal perceptions of digital risks, the importance of information and data security, as well as their view on the security culture within their respective organizations. This qualitative analysis, in addition to the analysis of Section A provided more context for the formulation of the general perception of the security posture already portrayed by the respondents and their respective organizations. From the analysis of this section, similar insights are gained regarding the overall level of security and awareness among the respondents and digitized organizations. The majority of respondent's show a good understanding of the term "digital risks" as well as of the importance of data and information security in modern organizations. Only two respondents out of 58 did not show interest in the importance of the topic. Most respondents who acknowledged the importance of the topic provided motivations that ranged between the repercussion from the lack of ample security, or the value gained from protection of sensitive information. With regards to the understanding of the term digital risks, the majority correctly associated digital risks to those introduced by conducting business in the digital environment, data being digitally used and stored, the usage of technology or any form of cybercrime. Only 7 respondents show unawareness of the term. The remainder of questions in this section focused on respondents' perspective of their organizations, with the aim to gain insight on the security culture, requirements and expectation within the workplace. It was found that 15 respondents do not receive adequate security training or education from their organizations. This raises a concern as assumed that training and education should form an integral part of any risk management plan. This high number of respondents who do not receive adequate security training could contribute to explaining why there is risky behavior found by several respondents' answers throughout the entire survey. With regards to perceived organizational security culture, it was found that the majority of respondents work in organizations that have a strict view on information and data security. Only in single cases did respondents mention a relaxed security environment.

From those respondents who indicated a relaxed security environment, one respondent stated that the security measures taken in the organization differ from person to person. This finding aids in the confirmation that although an organization may portray a strong security culture, the acts of one person can jeopardize or put the entire posture at risk. To investigate security culture in more depth, questions were asked regarding the required response to breach detection, what additional security measures are used to secure accounts, information system access and security, support group availability and accessibility within the organization. It was found that all the beforementioned organizational factors contribute positively to the general security posture of digitized

organizations. Additionally, it was found that regardless of the respondents' level of experienced stress in the workplace, stress does not appear as a determining factor of technology proficiency.

Overall and in general, adequate and responsive support groups are evident within organizations. Additionally, organizations do require employees to make use of multiple layers of security for the safeguarding of accounts. Additionally, they require employees to make timely escalations of security breaches and equip employees to safely establish networks connections. This is mostly done with the use of VPN's. This general conclusion is made through the consideration that the majority of all responses are positive in nature and promotes behavior that contributes to the strengthening of the security posture of respondents as well as their respective organizations. Although not explicitly stated, sufficient evidence exist that allows the assumption that organizations do implement risk management strategies. This assumption is made from the majority of respondents' answers that indicate strict security policies and requirements from their organizations. As a result, Q2 can thus be answered. It is also clear that organizations are aware of the impact and vulnerability of the workforce. This is seen through evidence captured by respondents that organizations do employ adequate measures to strengthen and prepare the workforce. Thereby, RQ1 is answered. The combined results from survey section A and B in conjunction with questions on security in Section C, the majority of respondents show high existing levels of security and threat awareness. RQ3 can thus be answered.

In conclusion of the analysis, it is clear that there exist a strong security culture within organizations and the reflection of such a security culture is visible within its employees. However, regardless of the level of security protocols in place within the organization, there will always be an employee who indulge in risky behavior. This can either be intentionally or unintentionally. Additionally, it is clear that observable environmental stimuli-response behavior is not adequate enough at forming an understanding of human behavior, thus answering RQ4. As such, proving the Behavioral Learning Theory to be less effective as proposed by Western Governors University (2020). The effectiveness of the Social Learning Theory to aid in the understanding of human behavior can thus now be brought into question.

5. Conclusion

It is clear that the understanding of human vulnerability and susceptibility to cyber threats are much deeper than an observable investigation. Human behavior is determined by a combination of both the understanding of personal psychological influences and the exposure to environmental stimuli, thus answering RQ4. The insight gained from this understanding provides great benefit to every aspect of an organization's security culture. Additionally, the human employee is at the forefront every facet of cyber security. It is crucial that organizations take every necessary step in order to prepare and promote security and risk awareness among the workforce. This trait is already visible within the workforce as the majority of respondents from an authentic dataset portray mindfulness towards cyber threats, thereby answering RQ3. With the rapid improvement and sophistication of modern cyber-attacks, it is impossible for organizations to completely prevent every attack. As attacks and breaches are inevitable, organizations should expect the worst and react accordingly, as it only takes one minor human mistake or error that leads to an organization's downfall. A strong sense of organizational awareness of the risk an impact of human vulnerability to cyber threats is visible. Organizations do in fact implement risk management strategies to prepare and improve the security posture of the workforce, thus answering RQ1 and RQ2. From the results there can be concluded that the impact of human vulnerability to cyber security is a risk that should not be overlooked or underestimated. The results from this paper emphasizes the importance of human-factor security in risk management and should receive priority of focus over the impact of technology itself.

6. Recommendation for future research

This study's investigation focuses solely on security awareness and behavior in relation to organizational information and data security. However, human behavior as a single factor is only one of many factors that defines the overall security posture of an organization. An organization's security posture and readiness to cyber threats cannot be derived only from the overall posture of the workforce. Due to the inevitability of a cyber-attack, the organization's ability to recover from such an attack holds significant weight. Sophos (2020) highlights that cybersecurity insurance aids in the recoverability of an organization after a ransomware attack. It was found that from all companies worldwide that had data encrypted from a ransomware attack, 94% of cases where the ransom was paid the cost was covered by insurance (Sophos, 2020). Additionally, Balbix (2021) highlights five factors that define an organization's security posture. These factors include existing protection controls and

processes, attack detection and containment, attack recoverability, the level of automation in the organization's security program and level of visibility in the organization's asset inventory. It is visible that the human employee has a role in each of these factors and needs to be investigated in order to determine an organization's security posture. It is thereby proposed that human behavior should be investigated in deeper detail that considers the human worker's role in the above-mentioned factors.

It is assumed that respondent's answers are subjective to their own perceptions of their security posture as well as their respective organization's posture. There exists the possibility that subjectivity in the answers could misrepresent reality. Thereby the question arises whether respondents' perception of their personal and their organizations security posture align with how security officials from the organization perceive it. It is thus proposed that in order to fully determine the security posture of an organization, additional investigation is needed with the focus on security officials in order to determine alignment of perceptions and expectations between organization and employee. The utilization of interviews as data collection methodology would allow the draw of more meaningful insight and improve the quality of data collection. This would prevent the collection of meaningless answers as seen with the usage of anonymous surveys. As it was concluded that The Behaviorist Learning Theory does not provide adequate support for the understanding of human behavior, further investigation with the use of The Social Learning Theory could be utilized to broaden this understanding.

References

- Ani, U. D., He, H., & Tiwari, A. (2019). Human factor security: evaluating the cybersecurity capacity of the industrial workforce. *Journal of Systems and Information Technology*, 21(1), 2–35. <https://doi.org/10.1108/JSIT-02-2018-0028>
- Ao, S. I., International Association of Engineers, WCECS (2008.10.22-24 San Francisco), & World Congress on Engineering and Computer Science (2008.10.22-24 San Francisco). (2008). *WCECS 2008, World Congress on Engineering and Computer Science, San Francisco, USA, 22-24 October, 2008*. IAENG International Association of Engineers.
- Balbix. (2021). *What is Security Posture?* <https://www.balbix.com/insights/what-is-cyber-security-posture/>
- Business Insider SA. (2020). *Beware of 'instant EFT' when buying online, regulators warn – you risk a great deal*. <https://www.businessinsider.co.za/reserve-bank-fsca-warns-against-instant-eft-for-online-shopping-2020-11>
- Evans, M., Maglaras, L. A., He, Y., & Janicke, H. (n.d.). *Human Behaviour as an aspect of Cyber Security Assurance*. GOVERNANCE & STANDARDS DIVISION. (2017). *IT Risk Management Framework DOCUMENT REVISION HISTORY*.
- Kure, H. I., Islam, S., & Razzaque, M. A. (2018). An integrated cyber security risk management approach for a cyber-physical system. *Applied Sciences (Switzerland)*, 8(6). <https://doi.org/10.3390/app8060898>
- Mcanyana, W., Brindley, C., & Seedat, Y. (2020). *INSIGHT INTO THE CYBERTHREAT LANDSCAPE IN SOUTH AFRICA*.
- Nobles, C. (2018). Botching Human Factors in Cybersecurity in Business Organizations. *HOLISTICA – Journal of Business and Public Administration*, 9(3), 71–88. <https://doi.org/10.2478/hjbpa-2018-0024>
- NRECA. (2011). *NRECA / Cooperative Research Network Smart Grid Demonstration Project Guide to Developing a Cyber Security and Risk Mitigation Plan*.
- Sophos. (2020). *THE STATE OF RANSOMWARE 2020*.
- Western Governors University. (2020). *What is the behavioral learning theory?* <https://www.plutora.com/blog/digital-risk>

Ensuring the Security of Space Systems from Eavesdropping Attacks

Caleb Richardson, Mark Reith and Wayne Henry

Air Force Institute of Technology (AFIT), Wright-Patterson AFB, USA

caleb.richardson@afit.edu

mark.reith@afit.edu

wayne.henry@afit.edu

Abstract: In a day and age where satellite communications are more important than ever to ensure global communication, establish international military power, and support our everyday way of life, satellite security must be at the forefront of innovation. However, eavesdropping attacks pose a serious threat to satellite communication systems that have not been adequately addressed. An eavesdropping attack threatens to put sensitive data in the wrong hands or even jeopardize critical missions. Research is needed to explore why defense against eavesdropping attacks is crucial, particularly for satellite systems. Three potential solutions to the problem are presented, addressing different challenges. Realistic solutions to the eavesdropping threat are needed urgently to defend the space domain from malicious threats.

Keywords: Satellite security, space systems, cyber security, eavesdropping attacks

1. Introduction

With space and cyber emerging as warfighting domains, anticipating the intersection of the two is vital. The most important issue to address in modern warfighting is ensuring the digital security of our nation's space systems. When it comes to satellites, we are concerned with three core security levels. First is the physical level. Physical security deals with securing the satellite's hardware and any related components on the ground. Next is the information level. This layer is concerned with ensuring that data sent to or from a satellite is confidential, maintains its integrity, and is available. Finally, there is the transmission layer, which focuses on ensuring the security of transmitting information to and from the satellite [10]. At the physical level, increased security may include more guards, stronger walls or classification levels.

This paper focuses on the information and transmission layers of security for satellites. One of the bigger problems that must be solved in regard to satellites is the eavesdropper attack [1]. This attack allows an unauthorized third party to intercept or potentially modify information being sent to or from a satellite. This type of attack has perpetuated itself in many forms across any device that sends information across a wire, the internet, or similar mediums. While this is not new to the technology world, it is one that is especially important with any satellite that transmits sensitive information. That is why finding a solution to this type of attack should take precedence over others. For the United States and its allies to stay at the forefront of technological capability, it is imperative to find a better solution to its satellites' most pressing security threat: the eavesdropper attack.

2. Background

Satellite communication operates similarly to communication across any network. The key layers of the Open Systems Interconnection (OSI) model (Physical, Data Link, Network, Transport, and Application) are still present [5]. However, the most significant difference is the distance that the radio signal has to travel.

While there are many types of satellites, this paper will primarily look at those in geostationary earth orbit (GEO). These satellites are the most used for communication because they are in a fixed position and can be reliably reached. Communications satellites are the most likely target of an eavesdropping attack because of the data that can potentially be acquired from them.

First, we briefly define an eavesdropping attack, and how it can be executed on a satellite. An eavesdropping attack, also known as a man-in-the-middle attack, is when an unauthorized agent listens in on communication they are not meant to receive. Due to the nature of how a satellite broadcasts a signal back down to earth, with the right equipment it is possible to intercept this communication without even being in close proximity to the satellite's ground receiver.

Satellites are not less prone to eavesdropping attacks because of new technologies. A potential eavesdropping attack has been modelled on the Thuraya network which implements the Geo Mobile Radio (GMR-1) standard

[2]. This attack was successfully theorized as a potential way to disrupt downlink communications. It could even be modified to modify the bit stream of the communication, manipulating the transmitted information [2]. Tests have even been conducted with equipment that an average person could purchase in which maritime communication was intercepted, including military and government communication [7, 8]. If this is possible to execute in an experimental environment, it is also a threat that could originate from someone with more malicious intentions.

3. Research

This research evaluates three potential solutions to the eavesdropper problem. Each of them presents a novel perspective on existing technologies that would make satellite transmission more secure against an eavesdropper attack. Each approach's fundamental structure is outlined followed by its strengths and weaknesses. After an overview of each solution, a composite solution is proposed using the most realistic and viable parts of each potential solution.

3.1 Quantum Key Distribution

One potential solution to the eavesdropping model could be implementing continuous variable quantum key distribution. This method uses quantum keys for encryption which greatly lowers the chance of an eavesdropping attack occurring [6]. The utilization of this encryption method allows for a constant secure key rate (SKR) under a lower aperture. This means that the downlink of the satellite will not have to transmit data at such a high frequency to the receiver. Having a lower transmission frequency decreases the area around the receiver where a potential eavesdropping attack could occur. Throughout the transmission process the SKR remains constant to ensure a reliable transmission [6].

In Figure 6 is an example of a transmitter A_a and a receiver A_b . A_{ex} represents the exclusion zone. In this exclusion zone, photons cannot be captured without the satellites knowing. Utilizing SKRs that are possible through quantum key distribution, the transmission frequency is low enough such that any potential eavesdropper will fall into one of two categories. They will either be too far away to intercept the transmission, or they will fall into the exclusion zone and the satellites will be notified of malicious activity and can act accordingly. Even if the eavesdropper attempts for optimal positioning, it will still fall into the exclusion zone [6].

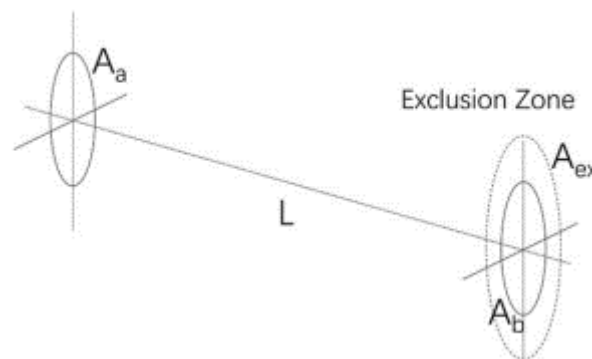


Figure 1: Satellite transmission between transmitter A_a and A_b [6]

The pros of this approach are that if it were implemented it would almost guarantee secure transmission against even the most sophisticated eavesdropper attacks. This technology is also faster than any current communication method so it would have massive implications in situations where information needs to be transmitted as close to real time as possible, such as when a decision needs to be made during a military conflict. The biggest downside to this solution is that the long-distance transmission of quantum keys is not yet at the point of reaching space as the furthest recorded transmission of a quantum key is currently 511 km [13]. Yet, as quantum key distribution technology grows in strength this will become an increasingly more attractive solution.

3.2 Cross Layer Architectures

Another option for increased satellite security is adopting a cross layer architecture (CLA). CLA is a design approach allowing non-adjacent layers of an OSI reference model to cooperate to optimize system performance. One solution would add additional protections in the TCP/IP stack as shown in Figure 2. While this approach was designed with streaming in mind it could also have applications in the cyber defense realm. The base approach

is as follows. In the transport layer, UDP-Lite would be used instead of UDP. This makes a distinction between sensitive and insensitive parts of a packet and any errors that occur within sensitive parts result in a packet drop. Because this is incompatible with the current IP security at the protocol level, a cross layer IP security (CL-IPsec) would be implemented. CL-IPsec is cross layer because it has access to the checksums of the UDP-Lite packet at the transport layer.

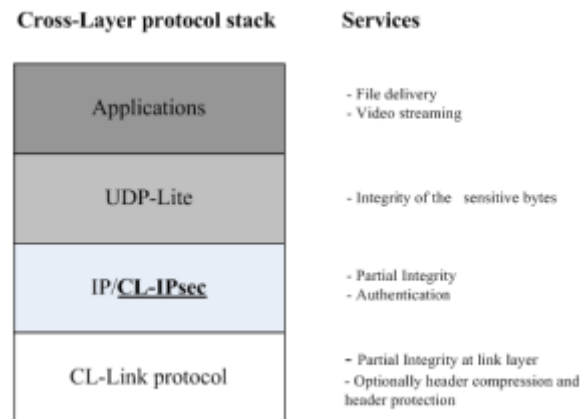


Figure 2: The CL-IPsec protocol stack [9]

The CL-IPsec network model is one that results in fewer packet drops of important data while gaining increased performance as it can forward packets to their destination as long as all the sensitive parts are untouched [9]. This could also have applications in security as a similar checksum could be used to ensure data is not tampered with via an outside attack. The CL-IPsec approach has both the benefit and detriment of modifying existing network protocol structures. Because the base protocols already exist, modifying them in this way would not be too difficult from the perspective of merely redefining what it expects from packets. This would make it a lot more difficult as we would have to figure out how to make a new protocol system accept both new and old types of packets.

3.3 Channel Estimation Error

Utilizing channel estimation error involves calculating what the estimated error of communication across a channel should be with and without third party eavesdropping and utilizing those values to determine whether the communication is being tapped [3]. The channel estimation error method is unique because it takes into account the possibility of there being multiple eavesdroppers in a given scenario and specifically tests the secrecy performance of the method with multiple eavesdroppers.

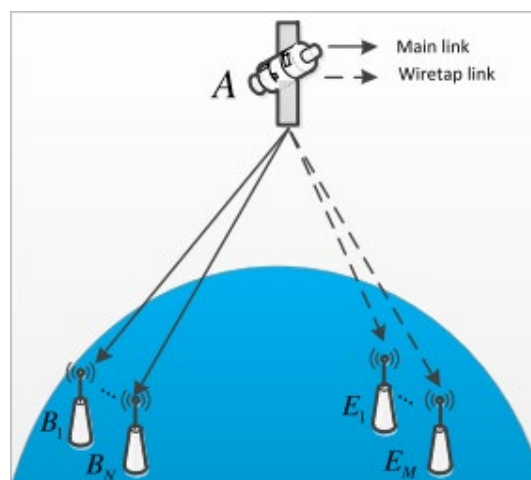


Figure 3: Model of a satellite transmission with both authorized (B_1 and B_N) and unauthorized (E_1 and E_M) listeners [3]

As mentioned previously, on satellite downlink it is possible for the eavesdropper to be in a separate geographical location and still intercept the communication. Because of this, it is possible for multiple

eavesdroppers to be in multiple locations and attempt to intercept or hijack a satellite downlink. The model for this attack shown in Figure 3 is comprised of a satellite with multiple authorized users (B_1 and B_N) and unauthorized eavesdroppers (E_1 and E_M). In the presence of channel estimation error, the secrecy performance on a satellite communication can be improved through a mathematical analysis of the satellite waves [3]. The upside of this approach is that it can theoretically work against multiple eavesdroppers at the same time. However, an implementation for it is not apparent and channel estimation error calculations can often be imperfect.

3.4 Proposed Solution

Based upon these solutions, the best way to combat eavesdropping on satellites would be to adopt a cross layer architecture that utilizes channel estimation error. While quantum key distribution is an enticing prospect for the future, it is the technology that is the furthest away from current adoption due to the current known limits of technology. It is important to keep it in the conversation so that when the technology does develop it can begin being applied to secure space transmission, however at the present time it is not feasible to include in this proposed solution. The concept of the cross-layer architecture is well designed and in order to increase transmission security revisiting the current network protocols will be imperative.

The most recent version of the "Overview for Space Communications Protocols" by the Consultative Committee for Space Data Systems (CCSDS) was released in 2014 [4]. This is a key player in the discussion of both governmental and quasi-governmental space system regulations and if this document has not been majorly revised in nearly a decade it is certainly worth revisiting and considering implementing additional security across the architecture layers. This could be done by adding checksums that check for two key bits of information. First, they would check to ensure all sensitive data in the packet is unaltered, similar to what the current model presents. Also, they could check to ensure the estimation error of the channel does not reach a level that might suggest tampering. Finding a way of adding channel estimation error calculations within the network stack could add even more security on top of the already suggested checksums. If tampering is detected the packet could not only drop, but temporarily drop all future packets to ensure data does not end up in the wrong hands. While there would be a delay before they are resent this would reduce the probability of the communication being intercepted.

These additions to the current communication protocols would certainly be impactful against the threat of eavesdropper attacks and would lay the groundwork for adoption of quantum key technology in the future that would give an even more sure-fire method of detecting unauthorized communication interference.

4. Conclusion

Eavesdropping attacks are a large threat to satellites and one that is not being properly addressed. It is imperative that this or a similar solutions be explored soon, because as satellites continue to play a larger role in communication they will become a ever looming target for adversaries. This research analyzed multiple models that could solve the problem. At the end of the day a proposed solution would revisit the current network protocols and implement additional checksums. These checksums would ensure that communications from a satellite reach their destination without being manipulated or intercepted.

Disclaimer

The views expressed are those of the authors and do not reflect the official policy or position of the US Air Force, Department of Defense or the US Government.

References

- Balakrishnan, S, Wang, P., Bhuyan A. and Sun, Z. 2019. "Modeling and Analysis of Eavesdropping Attack in 802.11ad mmWave Wireless Networks," in IEEE Access, vol. 7, pp. 70355-70370, doi: 10.1109/ACCESS.2019.2919674.
- Driessen, Benedikt. 2012. "Eavesdropping on Satellite Telecommunication Systems."
- Guo, K, An, K., Huang, Y. and Zhang, B. 2019. "Physical Layer Security of Multiuser Satellite Communication Systems With Channel Estimation Error and Multiple Eavesdroppers," in IEEE Access, vol. 7, pp. 96253-96262, doi: 10.1109/ACCESS.2019.2928751.
- "Overview of Space Communications Protocols," 2014, The Consultative Committee for Space Data Systems.
- Pan, Z. and Djordjevic, I. B. 2020. "Security of Satellite-Based CV-QKD under Realistic Assumptions," 2020 22nd International Conference on Transparent Optical Networks (ICTON), pp. 1-4, doi: 10.1109/ICTON51198.2020.9203397.

- Pavur, J., Moser, D., Strohmeier, M., Lenders V., and Martinovic, I. "A Tale of Sea and Sky: On the Security of Maritime VSAT Communications. doi:10.1109/SP40000.2020.00056
- Pavur, J.. 2020. "Whispers Among the Stars: A Practical Look at Perpetrating (and Preventing) Satellite Eavesdropping Attacks," blackhat USA 2021.
- Roseti, C., Luglio M., Provenzano S., Ramachandran, S. and Fairhurst, G. 2008. "A Cross-Layer Architecture for Satellite Network Security: CL-IPsec," 2008 4th Advanced Satellite Mobile Systems, pp. 82-87, doi: 10.1109/ASMS.2008.22. "Security Architecture for Space Data Systems." 2012, The Consultative Committee for Space Data Systems.
- Shea, S. 2020. "How to prevent network eavesdropping attacks". SearchSecurity. Retrieved November 6, 2021, from [https://searchsecurity.techtarget.com/answer/How-to-prevent-network-sniffing-and-eaves dropping](https://searchsecurity.techtarget.com/answer/How-to-prevent-network-sniffing-and-eaves-dropping).
- Slywczak, Richard. 2004. "Low-Earth-Orbit Satellite Internet Protocol Communications Concept and Design," National Aeronautics and Space Administration.
- Wang, C. and Lim C. 2021. "Long-distance quantum key distribution gets real," Springer Nature.
- Zheng, G., Arapoglou, P. and Ottersten B. 2012. "Physical Layer Security in Multibeam Satellite Systems," in IEEE Transactions on Wireless Communications, vol. 11, no. 2, pp. 852-863, doi: 10.1109/TWC.2011.120911.111460.

Work in Progress Papers

Mitigating Global Cyber Risk Through Bridging the National Incident Response Capacity Gap

Elisabeth Dubois and Unal Tatar

University at Albany, State University of New York, Albany, NY, USA

evdubois@albany.edu

utatar@albany.edu

Abstract: Cyber-attacks know no borders. Given the globally connected environment, no region or country is secure against cyber-attacks unless the entire world is secure or has cyber capabilities. Yet, whether preventive or reactive, cyber countermeasures require coordination and engagement of various organizations, government bodies, and citizens of different countries. Although a variety of countermeasures exist, Computer Security Incident Response Teams (CSIRTs) have been deemed necessary systems in defending against and preventing cyberattacks, further supporting a nation's cyber capacity and limiting the harm to citizens, businesses, and governments. Despite calls for establishing CSIRTs at the national level, especially toward protecting critical infrastructure and lives from cyber threats, various discrepancies exist based on a nation's resources, capabilities, and needs. Limited research delves into the cyber capabilities of low-income countries despite an emphasis on improving global cyber capacity, leading to a need to establish a framework for low-income countries to address the unique needs with lessons learned from existing standards. CSIRTs can improve the cybersecurity posture of countries, so we seek to investigate how low-income countries can better mitigate cyber threats through cyber capacity building, including the creation of CSIRTs. This work-in-progress paper aims to investigate cyber incident response capacity building at the national level in low-income countries and identify challenges they may face in contributing to a more secure global cyberspace. Stemming from this paper, we will conduct a survey of national CSIRTs in low-income countries and conduct semi-structured interviews to further investigate their role. The implications of our research are far-spreading, assisting academics, practitioners, and governments in developing research, processes, and policies to aid low-income countries in their national cyber capacity building.

Keywords: cyber incident response, CSIRT, cybersecurity, capacity building, low-income country

1. Introduction

Cyber-attacks know no borders. In the digitally connected world where over 4.66 billion of the world, or about 60% of the globe, are digitally connected via the Internet (Johnson, 2021), no region or country is secure against cyber-attacks unless the entire world is secure or has cyber capabilities. Much like the effect globalization has had on the transmission of COVID-19, all nations are indeed connected, and thus no one nation is secure unless all are secure, at least to some extent. In 2020, cybercrimes resulted in a 1 trillion-dollar loss to the global government, which made up more than 1% of the global GDP and signaled a 50% increase from 2018 (Smith et al., 2020). Worldwide, hundreds of thousands of cyber-attacks occur every day. Governments, businesses, critical infrastructure, and individuals face increased cyber risks, where cyber-physical risks to critical infrastructure are increasing exponentially. Think of the recent cyber-attack on Colonial Pipeline, which controls about half of the fuel supply along the East Coast of the United States. In fear that the attack had infected its systems, Colonial Pipeline turned them off, causing public panic with many buying at the gas pumps and spreading fear among the government that if not fixed soon, mass transit and factories would shut down as they would be unable to operate or distribute their products. Despite recovering, the ramifications of the pipeline attack are still being felt across the US. Concerns remain as to what will happen when a similar attack occurs in the US, other developed nations, or lower-income countries – will that country or the world be ready?

Most of the countermeasures to address cyber-attacks, whether preventive or reactive, require coordination and engagement of various organizations, government bodies, and citizens of different countries. At the forefront of these countermeasures are Computer Security Incident Response Teams (CSIRTs), which play a crucial role in national-level cyber security governance systems of countries, acting as the main coordinator and response team for cyber security incidents. CSIRTs can assume various other roles such as assisting in cyber security awareness and training, resolving incidents, cyber crisis planning and coordination activities, and/or recommending security notifications. In most countries, these teams initiate or promote the development of national-level policies and strategies. Thus, their roles often exceed basic incident coordination and awareness activities.

Globally, cybersecurity is viewed as not only a critical component to national and regional safety but a source of economic prosperity. National CSIRTs have been deemed active and necessary systems in defending against and

preventing cyberattacks and cybercrimes by supporting a nation's cyber capacity and limiting the harm to citizens, businesses, and governments. Although there are universal calls for the establishment of CSIRTs at the national level, especially toward protecting critical infrastructure and lives from cyber threats, various discrepancies exist based on a nation's resources, capabilities, and needs. Despite many countries having established a national CSIRT, generally, low-income countries often are left behind or face unique challenges affecting their effectiveness in countering or protecting their nation from cyber threats.

Limited research delves into the cyber capabilities of low-income countries despite an emphasis on improving global cyber capacity. This signifies a need to establish a framework for low-income countries to address the unique needs of low-income countries and create solutions to better mitigate cyber-risks. Low-income countries are of particular interest in this study because besides the common challenges in CSIRT creation and operation other more developed nations may face, low-income countries face a host of factors that are compounded by their socioeconomic state. For this study, developing or low-income countries are defined due to 'their levels of per capita income' (Sanford and Sandhu, 2003). Low-income countries are those that have an annual per capita income between \$875 – \$10,725, still lack critical infrastructure investments, have lower living standards, an undeveloped industrial base, and low Human Development Index (HDI) (IGI Global, 2020).

This work-in-progress paper aims to begin exploring the challenges of cyber incident response capacity building in low-income countries and propose a study to identify policy implications and capacity-building solutions to result in secure global cyberspace. The implications of our proposed research are expansive, offering academics and practitioners a foundation to further investigate and build cyber-capacity. In finding a solution to the cyber capacity-building challenges in low-income countries, the world will be more secure from cyber risks.

This paper will outline what others have done in this space and emphasize CSIRT capacity building, identify some of the challenges of CSIRT creation, and present the next steps in conducting a survey and series of interviews with CSIRT members in low-income countries to evaluate the role, challenges, and best practices of capacity building.

2. The Need for National CSIRTs

Effective mechanisms and structures at the national level are necessary to deal with cyber threats and build capacity (ITU, 2021). The absence of such structures or lack of strong national capacities poses serious risks to nations and the world. Thus, national CSIRTs can provide a solution. According to the European Union Agency for Cybersecurity (ENISA), risk management needs to support organizational improvement, performance, and decision making (ENISA, 2021). Within the past 5 years, CSIRTs Network was created to build the capacity of the information systems in Europe (EU) through improved cyber capabilities at the national level, greater EU-level cooperation and risk management, and incident reporting obligations (ENISA, 2016). Within this, ENISA stipulates that CSIRTs are essential to capacity building and managing risks, where the new network will develop confidence and trust between the Member States and promote more timely and effective operational cooperation (2016). Likewise, the International Telecommunications Union highlights the importance of CSIRTs and how the establishment and development of CSIRTs should be based on mature models, utilizing international collaborations, strong procedures, effective tools, and trainings (ITU, 2021).

3. Purpose of CSIRTs

Utilizing CSIRTs for capacity building can be an effective way to mitigate risks. The FIRST's CSIRT services framework classifies the services into five areas, including information security incident management, vulnerability management, information security event management, situational awareness, and knowledge transfer. These services are divided into reactive, proactive, and security quality management services (West-Brown et al., 2003). The services such as incident and vulnerability handling are presented as reactive. Proactive services range from intrusion detection to security assessments and maintenance of security infrastructures. The quality management service category covers the efforts related to training, awareness, consulting, risk analysis, and product evaluation.

Although CSIRTs work to improve many areas of cyber capacity, at the forefront of such capacity-building measures is the creation of a cybersecurity pipeline and human capabilities. The cyber security of any entity relies on the strength of its personnel and systems. As such, both developed and low-income countries should have CSIRT staff members that fulfill both the personal and technical skills to carry out cyber operations

(Carnegie Mellon University, 2021). In this, national CSIRTs should have enough technical experts, as well as experts trained in communication. According to the recent publication of the Carnegie Mellon University-Software Engineering Institute (CMU-SEI), skills like communication, presentation, diplomacy, the ability to follow policies and procedures, teamwork, critical thinking, problem-solving, time management, integrity, and coping with stress are commonly referenced as beneficial skills for CSIRT members, while technical skills are grouped into two categories, technical foundation and incident handling skills (Carnegie Mellon University, 2021).

CSIRTs, in many ways, build off the global emphasis on personnel training and improving employee soft and technical skills. Besides gaining skill knowledge, CSIRTs stipulate that certificates or incident response courses are essential features for CSIRT members (Carnegie Mellon University, 2021). Practitioners and researchers emphasize how cyber incident response capabilities or CSIRTs can be strengthened through education and training. One major emphasis is on cybersecurity curricula in academic and professional settings and the need to incorporate industrial control system education and training to increase the effectiveness of incident response efforts of CSIRTs and government agencies (Marzitelli et al., 2015). Within this, CREST, an international not-for-profit organization (as well as a series of other organizations), helps professionals working towards qualifications or who would like to acquire, maintain, or extend their knowledge and build their capability (CREST, 2019).

Many CSIRTs offer lists of tools and processes aimed at helping build capacity or improve the effectiveness of CSIRTs. The National Cyber Security Centre in the Netherlands guide, as well as several other CSIRT guides, discusses the importance of appropriate tools and how the lack of sufficient tools can cause inefficient work and greater risk. Today, automated tools are one of the most critical weapons in a CSIRTs toolbox, as incident loads increase and cyber threats continue to grow (Van der Kleij et al., 2017).

Similarly, a host of national and international organizations and governments offer a variety of tools and trainings to assist in capacity building. In a literature review by Chowdhury & Gkioulos (2021), they discuss that given the cyber threats to the national CI, various sectors have instilled trainings and testbeds to better preparedness against cyber threats. They found that cybersecurity training towards national security is split between awareness, technical, specialized CS, and incident response and recovery training (Chowdhury and Gkioulos, 2021). The Global Cyber Alliance (GCA) has developed several free cybersecurity tools and processes to meet the needs of CSIRTs and build cyber capacity. Within this, GCA provides several free toolkits to improve the cybersecurity and integrity of small businesses, elections, journalists, and financial institutions (Global Cyber Alliance, 2021). These toolkits vary in use and offerings, but they entail steps to not only improve the cyber hygiene of the particular party but provide tiered tools and recommendations for strong cyber protection across the board. Outside of these toolkits, GCA also offers a variety of useful tools to limit cyber risks via mobile and IoT devices and web searching, including Domain-based Message Authentication, Reporting, and Conformance (DMARC), Automated IoT Defence Ecosystem (AIDE), and Quad9. All of the tools provided by GCA are created in collaboration with other organizations, government entities, and individuals around the world to better address the challenges of cyber hygiene, response, and capacity building. Additionally, the Center for Internet Security publishes security guidelines, checklists, and tools for different platforms, operating systems, and other IT systems (Center for Internet Security, 2021). FIRST provides a list of resources for technical and procedural guidelines and hosts trainings on threat intelligence (FIRST, 2021). The International Telecommunications Union (ITU) also provides information about CSIRT tools as well as many procedural and training opportunities (International Telecommunication Union, 2021). Organizations and CSIRTs rely on SANS for free resources and tools to improve or build cyber security capabilities. The SANS Faculty Free Tools, alongside other resources, provide open source tools to support the work of organizations or governments to promote better security (SANS, 2021a). Within this, SANS also offers a number of quick reference guides and cheat sheets for IT security, digital forensics, incident response, offensive operations, cloud security, industrial control systems (ICS), and all-around defense (SANS, 2021b). A host of other tools are needed and utilized to store and track incidents, share cyber threat intelligence, coordinate data sharing and collaboration, as well as cater to one or more of the CSIRT services (including penetration testing, web application security).

4. The Need for National CSIRTs in Low-Income Countries

Although the need and purpose of national CSIRTs have been established, there are critical factors that make them necessary for low-income countries. As previously mentioned, low-income countries are those that have an annual per capita income between \$875 – \$10,725, still lack critical infrastructure investments, have lower

living standards, an undeveloped industrial base, and/or low Human Development Index (HDI) (IGI Global, 2020). Such countries include Nigeria, Indonesia, Bangladesh, Zambia, Cote d'Ivoire, Trinidad and Tobago, Togo, Malawi, etc.

At the forefront of the need for national CSIRTs among low-income countries is if low-income countries are not secure or do not have cyber capabilities, then the entire world faces greater risks. Yet, for low-income countries, more significant challenges exist in establishing a CSIRT. According to Grobler and Bryk, several challenges exist during the establishment of a CSIRT. Said challenges are clarifying mandates and policy issues, obtaining management support, securing funding, availability of staff and other resources, understanding cultural and governmental differences, training staff, and developing operational and technical procedures (Grobler and Bryk, 2010). Of these challenges CSIRT's ability to secure the relevant funds is vital. However, less-developed countries have fewer funding alternatives when compared to developed ones (Skierka and Hohmann, 2015). For these countries, some alternative options include government sponsorship, fee-based services, in-kind supports, or a combination of these could be funding alternatives for national CSIRTs (Jalal et al., 2018). Likewise, in low-income countries, the availability of qualified staff or solid technology pipeline, training options, external support, and cultural differences create greater challenges due to limited critical infrastructure, lower living standards, underdeveloped industrial base, and low HDI. Through capacity-building support with international collaborations, affordable training opportunities, affordable tools for reliable infrastructure, and a greater focus on workforce improvement, low-income countries operating CSIRTs can better address the common challenges.

5. Research Objectives

This study will explore the challenges of cyber incident response capacity building at the national level for low-income countries. The aims and objectives of this research are:

- **Obj1:** Identify the current CSIRT state in low-income countries (e.g., services delivered, personnel capacity, technical capability, technological infrastructure, legal base, and whenever applicable, identify the level of cybersecurity maturity for each of the countries, including governance structure)
- **Obj2:** Identify challenges of establishing and maintaining a national CSIRT
- **Obj3:** Identify solutions to mitigate the challenges of CSIRTs and enhance capacity building in low-income countries.

6. Next Steps: Methodology and Expected Findings

Following an expanded desk review of academic and grey literature (e.g., reports of security vendors, independent organizations, government entities, national CSIRTs) to list the national CSIRT services and identify the relevant organizational models, applied incident handling processes, workflows, required human skill sets, training resources, applicable toolsets, and best practices in capacity development we plan to collect survey data and conduct interviews. A geographically diverse quantitative survey will be used to investigate the challenges of cyber incident response capacity building at the national level among low-income countries. We will aim for a representative sample of members of low-income countries CSIRTs or national cybersecurity teams. Following the survey, we plan to conduct select interviews with a handful of the survey respondents to identify challenges and best practices from a consultant's perspective and learn more about the available resources for low-income countries to develop CSIRT services.

Once completed, this study seeks to present solutions to mitigating cyber risks faced by low-income countries. Based on the desk review, such solutions may include capacity-building support through international collaborations, affordable training opportunities, affordable tools for reliable infrastructure, and a greater focus on workforce creation and a technology pipeline.

References

- Carnegie Mellon University, 2021. Cyber Workforce Development [WWW Document]. Carnegie Mellon Univ. Softw. Eng. Inst. URL <https://www.sei.cmu.edu/our-work/cyber-workforce-development/index.cfm> (accessed 12.5.21).
- Center for Internet Security, 2021. Center for Internet Security [WWW Document]. Cent. Internet Secur. URL <https://www.cisecurity.org/> (accessed 12.5.21).
- Chowdhury, N., Gkioulos, V., 2021. Cyber security training for critical infrastructure protection: A literature review. *Comput. Sci. Rev.* 40, 100361. <https://doi.org/10.1016/j.cosrev.2021.100361>
- CREST, 2019. CREST Representing the Technical Information Security Industry.
- ENISA, 2021. Risk Management [WWW Document]. Eur. Union Agency Cybersecurity ENISA. URL <https://www.enisa.europa.eu/topics/threat-risk-management/risk-management> (accessed 12.5.21).

- ENISA, 2016. CSIRTs Network [WWW Document]. Eur. Union Agency Cybersecurity ENISA. URL <https://www.enisa.europa.eu/topics/csirts-in-europe/csirts-network> (accessed 12.5.21).
- FIRST, 2021. FIRST Security Reference Index [WWW Document]. FIRST — Forum Incid. Response Secur. Teams. URL <https://www.first.org/resources/guides/reference> (accessed 4.26.21).
- Global Cyber Alliance, 2021. Global Cyber Alliance: Enabling a Secure and Trustworthy Internet [WWW Document]. GCA Glob. Cyber Alliance Work. Eradicate Cyber Risk. URL <https://www.globalcyberalliance.org/> (accessed 3.31.21).
- Grobler, M., Bryk, H., 2010. Common challenges faced during the establishment of a CSIRT, in: 2010 Information Security for South Africa. Presented at the 2010 Information Security for South Africa (ISSA), IEEE, Johannesburg, South Africa, pp. 1–6. <https://doi.org/10.1109/ISSA.2010.5588307>
- IGI Global, 2020. What is Developing Countries | IGI Global [WWW Document]. URL <https://www.igi-global.com/dictionary/cyber-capability-framework/7401> (accessed 12.15.20).
- International Telecommunication Union, 2021. ITU Cybersecurity Activities [WWW Document]. ITU. URL <https://www.itu.int:443/en/action/cybersecurity/Pages/default.aspx> (accessed 12.5.21).
- ITU, 2021. National CIRT [WWW Document]. ITU. URL <https://www.itu.int:443/en/ITU-D/Cybersecurity/Pages/national-CIRT.aspx> (accessed 12.5.21).
- Jalal, I., Yusof, M., Shukur, Z., Mokhtar, M., 2018. A Model for Afghanistan's Cyber Security Incident Response Team. *Int. J. Adv. Sci. Eng. Inf. Technol.* 8, 2620–2626. <https://doi.org/10.18517/ijaseit.8.6.6692>
- Johnson, J., 2021. Internet users in the world 2021 [WWW Document]. Statista. URL <https://www.statista.com/statistics/617136/digital-population-worldwide/> (accessed 8.9.21).
- Marzitelli, R., Monroe, K., Newmeyer, K., Parker, C., Pheils, D., Shih, L., Silas, M.A., Tu, M., 2015. *National Cybersecurity Institute Journal* 2, 80.
- Sanford, J.E., Sandhu, A., 2003. *Developing Countries: Definitions, Concepts and Comparisons*. Nova Publishers.
- SANS, 2021a. SANS Faculty Free Tools.
- SANS, 2021b. The Ultimate List of SANS Cheat Sheets. Inst. URL <https://www.sans.org/blog/the-ultimate-list-of-sans-cheat-sheets/> (accessed 12.5.21).
- Skierka, I., Hohmann, M., 2015. CSIRT Basics for Policy-Makers.
- Smith, Z.M., Lostri, E., Lewis, J.A., 2020. The Hidden Costs of Cybercrime. McAfee.
- Van der Kleij, R., Kleinhuis, G., Young, H., 2017. Computer Security Incident Response Team Effectiveness: A Needs Assessment. *Front. Psychol.* 8, 2179. <https://doi.org/10.3389/fpsyg.2017.02179>
- West-Brown, M.J., Stikvoort, D., Kossakowski, K.-P., Killcrece, G., Ruefle, R., Zajicek, M., 2003. Handbook for Computer Security Incident Response Teams (CSIRTs) 223.

An Exploration on Advanced Persistent Threats in Biocybersecurity and Cyberbiosecurity

Xavier-Lewis Palmer^{1,2}, Lucas Potter¹ and Saltuk Karahan²

¹Department of Biomedical Engineering, Old Dominion University, Norfolk, VA, USA

²School of Cybersecurity, Old Dominion University Norfolk, VA, USA

xpalm001@odu.edu

lpott005@odu.edu

skarahan@odu.edu

Abstract: Novel and complex digital threats that are increasingly interwoven with means and products of biology that can affect society. Much work in Biocybersecurity/Cyberbiosecurity (BCS/CBS) discuss vulnerabilities, but few deeply address malicious actor varieties as attacks at this intersection are new. The path to those attacks remains mostly theoretical, presenting considerable difficulty to accomplish in practical scenarios. In terms of advanced persistent threats (APTs) this of course needs to change as biomanufacturing facilities are at risk, considering Covid-19 and other potential pandemics. Further attacks are not out of reach and thus we must start to imagine how BCS APTs may appear. This paper in progress aims to open discussion regarding the definition of the concept BCS/CBS APTs and their implications, as well as create call to action for increased attention.

Keywords: Biocybersecurity, Cyberbiosecurity, Covid-19, Cybersecurity, Biosecurity, Biomanufacturing,

1. Introduction: BCS/CBS and Security Studies

Cybersecurity has already found its place within security studies and is explicitly stated in several national security documents with an increasing level of emphasis for the past two decades (Berger and Roderick, 2014). The most recent Interim National Security Strategic Guidance of the Biden administration lists cyber and digital threats along with pandemics and biological risks after stating that “recent events show all too clearly that many of the biggest threats we face respect no borders or walls, and must be met with collective action.” (Biden, 2021). It is also stated in the same document that cybersecurity is to be made “a top priority, strengthening our capability, readiness, and resilience in cyberspace”.

As the cyber domain has an increasing presence in security studies, new subdomains gain significance as well. Social Cybersecurity defined as “an emerging scientific area focused on the science to characterize, understand, and forecast cyber-mediated changes in human behaviour, social, cultural, and political outcomes” has been seen as an emerging subdomain (Carley et al, 2018). Similarly, biocybersecurity is another such subdomain which has the potential to influence our understanding of warfare with its unique nature encompassing several aspects of biology and information technology. While there is still not a clear explanation for the rare use of biological warfare throughout history, the risk of attribution and only a few state actors’ access to the technology (considering these were accountable actors) were among the possible explanations for this rare use (Carus, 2017). However, the advancement in biotechnology has the potential to change this fact when the effects of biocybersecurity is considered. This fact has become even more prominent as we started to see the use of biological components as a target by APTs. APTs are often associated with state actors who can fund such sophisticated attacks and could have long term political goals to be gained from these attacks. The trend in the conduct of APTs further reinforces the necessity to reconsider BCS/CBS as influential subdomains within cybersecurity. While early attempts within bio-cyber APTs were limited to the disinformation campaigns, the recent trends include interventions into biotechnological systems and the conduct of these interventions at scale is looming in the horizon. Furthermore, the access to the biocyber attack capabilities by rogue states and terrorist organizations has significantly increased due to the nature of all cyber-attack capabilities. In the next section, we will look further into the world of APTs in biocybersecurity.

2. BCS/CBS and APTs

Moves of APTs are rather easy on the behalf of state actors because they have the resources to supply for multiple, well-outfitted, and long-lasting teams. Establishing control can take place through the deployment of a combination of moves in social media or bioeconomic systems from social engineering. From there, teams can establish their place within a system and use it as a launching point. Well placed malware, social media campaigns, and or employees can be used to deepen presence and permanence in society. Eventually, the APT

finds a groove in a consistent system for examining the network, extracting or attacking assets, and keeping its footing more securely. However, some APTs can re-attack in coming in the form of a different group or at a different time, with better resources. The persistent nature cannot be understated. This is a general workflow and is something to be contended with by more of the bioeconomy, and this became apparent in 2021 with the finding of Tardigrade, a metamorphic, semi-autonomous APT used to target biomanufacturing facilities as revealed by Bioeconomy Information Sharing and Analysis Center (Newman, 2021).

3. Insights Taken from Leading Experts Across Academia, Government, and the Private Sector

In addition to Peccoud et al's (2018) and Murch et al's (2018) initiation and spurring of the conversations on the blurred territory of cybersecurity, biosecurity, and cyber-physical security, Reed and Dunaway (2019) delivered a further exploration of labs of the future, finding that governments still face significant challenges in preparation for the unique threats that Cyberbiosecurity presents. This is in agreement with George (2019) who reflected on posed national security implications as these core technologies continue to converge in use. This concern is amplified through findings of Millett et al (2019) who explored perceptions of risk in the Biotech sector through a study which surveyed anonymized leaders in biotechnology and biosecurity, concerning perceived risks. Their work revealed that a considerable number of the participants had difficulty assessing risks to their organizations in the face of cyber-related threats. Some experts surveyed were unaware of dedicated resources that could be used to mitigate risks in biotechnology, and many believed that more time and resources are needed to be dedicated. A broader picture revealed that there are considerably wide gaps for malicious actors to take advantage of. As the poll was online and limited, further research is necessary; however, concerns reliably exist among the question of who is most unprepared. This altogether, indicates need for a closer look at the state of biomanufacturing and need for surgical reassessments of infrastructure.

Mueller (2020) discussed at length challenges that COVID--19 introduced, highlighting bioeconomy diversity and that they contain unique elements that require further assessment. It's important to note that the networks and interactions of components of vital elements such as agriculture, animal husbandry, restaurants, pharmacies, and processing plants will differ by locale and subjected policies. These are each further open to compromise in part due to varying pressures that pandemics impose on the relevant populations. Mueller gave recommendations on responding to cyber-bio convergences, which benefit from cross-discipline insights; follow through on these recommendations are vital given that allowed incomplete public awareness of relevant cyber-bio convergent potential and their potential routes to affect key public institutions and businesses, combined with wide gaps in expertise across fields that comprise BCS /CBS can sustain vulnerabilities that can be exploited by malicious actors. This is especially critical in the ongoing management of Covid-19, considering shifting vaccine availability and heightened importance of the facility capacity available to administer care. Jordan et al (2020) expanded security conversation in this space through artificial intelligence as bioeconomies are open to novel AI/ML threats. When considering AI, one needs to consider competing ethics at play within computing, and biological data. Threats using cyber and bio databases, along with focus on emergent functions require even more careful analysis as grey areas using such research are many and the impact of the threat is not easy to see. Jordan et al (2020) lists low to high-risk national security phenomena relating to data, using demonstrated and possible cases of exploits, which show a wide field open to defend. Proper application of AI algorithms can boost efforts of detecting CBS/BCS threats, but also improve the quality of malware created, morphing and accelerating the BCS/CBS landscape. This altogether enhances APT risks for biomanufacturers and as actors apply this technology with increasing efficiency. Tardigrade may just be only the APT successfully detected, with many others out there potentially in development. Questions of when biocomputing aided APTs may emerge within biological systems are valid as relevant technique and capacity become increasingly available.

4. Proposed Approach to BCS/CBS and Concluding Thoughts

Starting from the understanding of biocyber threat, there are several measures that can be taken to secure ourselves from threats within BCS/CBS subdomain. These measures extend from BCS/CBS advocacy to further study of means of detection and building capacity (and resilient infrastructure) within this subdomain. Examining the security implications of big data in life sciences, "vulnerabilities of datasets to cyber intrusion and design of biological weapons derived from the integration and analysis of Big Data in the life sciences" has been listed as one of the risks associated (Berger and Roderick, 2014). Considering the fact that "computers can be compromised by encoding mal-ware in DNA sequences, and bio-logical threats can be synthesized using publicly available data" as stated by Peccoud et al., there needs to be further research in policy guidelines and detection

mechanisms (Peccoud et al, 2018). Considering the intrinsic link between biotechnology and computing, advanced infrastructure investment and capacity building remain necessary steps. There have already been recommendations for this including incentivizing biotech companies, cyber-centers for disease control and prevention, mandating gene guard systems, and consumer education (Pauwels and Vidyarthi, 2016). While the foundational contribution of existing approaches in disciplines linked to biocybersecurity are recognized, their incompetency in capturing particularly “the emergent features related to biological and biomedical systems” necessitates a genuine approach (Schabacker et al, 2019). Further infrastructure development should include education and reassessment that focuses knowledge on both modern and emergent cyber-bio gaps and policies in terms of incentives, along with updated regulation in infrastructure designs and deployment, where most critical to national security. Beyond these measures at the detection or infrastructure or regulation levels, a stronger and wider awareness campaign highlighting increased cyber-bio convergence is needed regarding BCS/CBS subdomain when the increasing potential APT attacks to occur in this field are considered. This awareness can only be accomplished with further advocacy in the recognition of the field as a significant subdomain of cybersecurity, increasing discussion of vulnerabilities, threats and policy topics in academic circles and government entities.

5. Future Work

2018 National Biodefense Strategy mentions the word “cyber” only twice throughout the 30 page document. Those two referrals are in the contexts of including cybersecurity in “strengthening biosafety and biosecurity practices” and “ensuring the safety, security, and reliability of supply chains to provide necessary medical countermeasures and other materiel support during bioincidents”. The authors argue that considering the increasing ransomware like attacks at medical facilities, increasing reliance on technology and increasing expectation for genetic screening to be part of preventative care, cybersecurity should be considered and essential component of biosecurity or biodefense and this should be reflected in strategy documents.

Department of Health and Human Services has the leading role in the governance of biosecurity and Presidential Memorandum on the Support for National Biodefense (NSPM-14) accompanied the issuance of the National Biodefense Strategy. As indicated in the National Biodefense Strategy, NSPM -14 “creates a dedicated mechanism to coordinate federal biodefense activities and assess the effectiveness with which the National Biodefense Strategy’s goals and objectives are being met.” While establishment of a dedicated mechanism is very promising, the increasing link between cybersecurity, biosecurity and national security requires a clear and stronger multi-agency involvement and a coordinated effort in building strategy. With its national security, economic and political dimensions, and the increasing sophistication of biocyber attacks, the sources of the attacks weigh towards intentional attacks from nation states rather than naturally occurring and accidental outbreaks. The future work in this topic would include a survey of existing mechanisms for an effective coordination and cooperation, determination of the gaps in strategy and policy and recommendations on how to approach the biocybersecurity problem.

References

- Berger, K. M., & Roderick, J. (2014). National and transnational security implications of big data in the life sciences. Washington, DC: American Association for the Advancement of Science.
- Biden Jr, J. R. (2021). Interim National Security Strategic Guidance. Executive Office of the President, Washington DC.
- Carus, W. S. (2017). A short history of biological warfare: from pre-history to the 21st century (Vol. 12). Government Printing Office.
- George, A. M. (2019). The national security implications of cyberbiosecurity. *Frontiers in bioengineering and biotechnology*, 7, 51.
- Jordan, S. B., Fenn, S. L., & Shannon, B. B. (2020). Transparency as Threat at the Intersection of Artificial Intelligence and Cyberbiosecurity. *Computer*, 53(10), 59-68.
- Karahan, S., Wu, H., & Armistead, L. (2019). Evolution of US cybersecurity strategy. In *International Conference on Cyber Warfare and Security* (pp. 168-176). Academic Conferences International Limited.
- Kathleen M. Carley et al. (2018) “Social Cyber-Security,” in *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP- BRiMS 2018, Washington, DC, USA, July 10–13, 2018, Proceedings*, ed. Halil Bisgin et al. (New York: Springer, 2018), 389–94.
- Millett, K., Dos Santos, E., & Millett, P. D. (2019). Cyber-biosecurity risk perceptions in the biotech sector. *Frontiers in bioengineering and biotechnology*, 7, 136.
- Mueller, S. (2020). Facing the 2020 Pandemic: What does Cyberbiosecurity want us to know to safeguard the future?. *Biosafety and Health*.

- Murch, R. S., So, W. K., Buchholz, W. G., Raman, S., & Peccoud, J. (2018). Cyberbiosecurity: an emerging new discipline to help safeguard the bioeconomy. *Frontiers in bioengineering and biotechnology*, 6, 39.
- Newman, L.H., (2021) Devious 'Tardigrade' malware hits biomanufacturing facilities. *Wired*. Available at: <https://www.wired.com/story/tardigrade-malware-biomanufacturing/> [Accessed November 25, 2021].
- Pauwels, E., & Vidyarthi, A. (2016). How Our Unhealthy Cybersecurity Infrastructure is Hurting Biotechnology. *Wilson Briefs*. Washington, DC: The Wilson Center.
- Peccoud, J., Gallegos, J. E., Murch, R., Buchholz, W. G., & Raman, S. (2018). Cyberbiosecurity: from naive trust to risk awareness. *Trends in biotechnology*, 36(1), 4-7.
- Reed, J. C., & Dunaway, N. (2019). Cyberbiosecurity Implications for the Laboratory of the Future. *Frontiers in bioengineering and biotechnology*, 7, 182.
- Schabacker, D. S., Levy, L. A., Evans, N. J., Fowler, J. M., & Dickey, E. A. (2019). Assessing cyberbiosecurity vulnerabilities and infrastructure resilience. *Frontiers in bioengineering and biotechnology*, 7, 61.

Late Submissions

Cyberwarfare and its Effects on Critical Infrastructure

Humairaa Yacoob Bhaiyat and Siphesihle Philezwini Sithungu

Academy of Computer Science and Software Engineering, Faculty of Science, University of Johannesburg, South Africa

humairaabhaiyat@gmail.com

siphesihles@uj.ac.za

Abstract: The growth and capabilities of cyberspace have brought about many advantages to societies. Individuals and businesses have used cyberspace for easier communication, but nation-states also utilise it to improve the functioning of their critical infrastructure. Critical infrastructures provide vital services such as the health, safety and security needed for the efficient functioning of societies. However, vulnerabilities in cyberspace have made cyberattacks such as cyberwarfare possible. Cyberwarfare is an international concern due to the negative impact it can have on critical infrastructure. This paper aims to discuss cyberwarfare and the potential effects that it can have on critical infrastructure. This paper follows a theoretical research methodology to provide an understanding of cyberwarfare. In addition, the paper provides a better understanding of the impact that cyberwarfare can have on critical infrastructure. The paper contains an exhaustive definition of cyberwarfare. Since cyberwarfare is a type of cyberattack, it is similar but not the same as other cyberattacks such as cybercrime and cyberterrorism. Therefore, to gain a clear understanding of cyberwarfare, the paper discusses cyberwarfare, cybercrime, and cyberterrorism. The paper also discusses some of the most significant cyberwarfare incidents. Since the effects can be devastating, critical infrastructure must be protected from cyberwarfare. A survey of techniques for protecting critical infrastructure from cyberwarfare is presented. The identified incidents highlight the effects that cyberwarfare can cause. Hence, the possible effects that cyberwarfare can cause on critical infrastructure is discussed. Due to the negative effects of cyberwarfare, nations need to be prepared to protect their critical infrastructure from cyberwarfare. Therefore, the paper also discusses the authors' stance on South Africa preparedness to defend themselves in the event of cyberwarfare.

Keywords: cyberattack, cyberwarfare, critical infrastructure, critical information infrastructure, effects, SCADA systems

1. Introduction

Developments in Information and Communication Technologies (ICT) has led to adverse consequences within cyberspace (Izycki & Vianna, 2021). Cyberwarfare has utilised cyberspace to conduct attacks whose effects are felt both in cyberspace and the physical world (Almeida, Doneda & de Souza Abreu, 2017). The effects of cyberwarfare are highly detrimental to the Critical Infrastructure (CI) of a nation. Incidents of cyberwarfare have already occurred in states such as Russia, Iran, and the United States. Therefore, techniques to protect against such cyberwarfare incidents are vital. Countries such as Russia, Iran, North Korea, and Israel have already started preparing to protect themselves from such incidents (Colarik and Janczewski, 2012). But are countries in Africa, such as South Africa, prepared to defend themselves against cyberwarfare?

This research paper aims to discuss the concept of cyberwarfare and the possible effects it can have on CI. The paper is structured in the following manner to achieve this objective: Section 2 explains what cyberwarfare is. Section 3 provides a discussion on cybercrime, cyberterrorism, and cyberwarfare. Section 4 provides a discussion of the different cyberwarfare incidents that occurred worldwide. Section 5 provides a few techniques that can be used to protect CI from cyberwarfare. Section 6 highlights the possible effects of cyberwarfare on CI, and section 7 discusses the authors' stance if South Africa is prepared to defend itself against cyberwarfare. The last section, section 8, concludes the paper.

2. What is Cyberwarfare

Currently, there are several definitions of cyberwarfare. Clarke and Knake (2010) describe cyberwarfare as actions where a nation-state will penetrate another nation's networks or systems to cause damage or disruption. This definition implies that the target and source of cyberwarfare actions are only between nation-states to advance a national agenda such as information superiority. However, Robinson, Jones and Janicke (2015) describe that cyberwarfare includes state actors and non-state actors such as businesses or hackers whose aim is not a national agenda. Another definition described by Cornish, Livingstone, Clemente, and Yorke (2010) is that cyberwarfare is a conflict between states, but it can sometimes involve non-state actors. The target of these attacks can be industrial, military, civilian targets or a server room hosting several client information. This definition implies that cyberwarfare targets can be different infrastructures, including state and non-state actors. Cyberwarfare includes network attacks and special and technical operations and defence (Parks &

Duggan, 2011). It usually involves actions that can cause political or military effects (Almeida, Doneda & de Souza Abreu, 2017).

Cyberwarfare uses cyberspace to cause cascading kinetic effects (Almeida, Doneda & de Souza Abreu, 2017). Kinetic effects are actions that can affect something or someone in the real world (Parks & Duggan, 2011). These kinetic effects can create results that either cause or contribute to a severe threat to the security of a nation or result in actions taken to respond to such a threat (Almeida, Doneda & de Souza Abreu, 2017). Cyberspace is a virtual space that uses electronics or electromagnetic spectrum to store, modify, exchange, and create information using ICT (Mueller, 2020; Robinson, Jones & Janicke, 2015). Cyberwarfare actions include disruptions to CI and information conflict that can lead to material and social disruptions (Parks & Duggan, 2011). Similar cyberattacks that can also lead to disruptions is cybercrime and cyberterrorism. The following section discusses cybercrime, cyberterrorism, and cyberwarfare to understand what they are and how they differ from one another.

3. Cybercrime, Cyberterrorism and Cyberwarfare

Section 3.1 addresses cybercrime, section 3.2 explains cyberterrorism, and section 3.3 looks at cyberwarfare.

3.1 Cybercrime

Chandra and Snowe (2020) describe cybercrime as an action that uses computer technology to commit a crime. Cybercrime can be classified into two groups, “cyber-enabled” crime and “cyber-dependent” crime (McGuire & Dowling, 2013). Cyber-enabled crimes existed before but are now easier to pursue due to computer technology. Examples are white-collar crime, identity theft, drug trafficking, stalking, etc. Cyber-dependent crimes are crimes that would not exist without cyber technology. These crimes use malware to perform criminal acts (Sarre, Lau & Chang, 2018).

Cybercrime has been growing due to the number of financial benefits. It is easy to obtain money and data due to many users in cyberspace with limited knowledge of how technology works (Bernik, 2014). Cybercrime, cyberterrorism, and cyberwarfare all fall under the same category: cyberattacks (Gazula, 2017). But what makes cybercrime different from cyberterrorism and cyberwarfare is the attacker’s intention. Cybercrime is often committed for personal reasons such as personal gain or the desire to harm others physically or psychologically (Brenner, 2006). An example of an incident of cybercrime occurred during January 2010 and August 2011 when a 23-year-old man, Edward Pearson, stole 200,000 Paypal account details, 2,700 bank card numbers, and 8 million identities using the malware Zeus and SpyEye. This is an example of cybercrime with the intention of personal gain (Broadhurst, Grabosky, Alazab, Bouhours, & Chon, 2014).

3.2 Cyberterrorism

Cyberterrorism describes acts that use computer technology to intimidate a civilian population, influence the policy of a government, or influence the conduct of the government through intimidation (Robinson, Jones & Janicke 2015). Brenner (2006) states that cyberterrorism is different from cybercrime because cyberterrorism attacks are often due to political reasons rather than personal reasons.

Cyberterrorism and cyberwarfare have similar characteristics, goals, and manner of execution, but the difference between these two cyberattacks is their strategy, targets, and the perpetrators. Cyberterrorism is believed not to be sponsored by a state, while cyberwarfare is often sponsored by a state (Vilić, 2017). The strategy of cyberterrorism involves acts of violence through mass destruction, assassination, or kidnapping (Robinson, Jones & Janicke, 2015). A cyberattack that targets networks and computers is believed to be an act of cyberterrorism when the effect of the attack causes destruction enough to produce fear that can be comparable to a physical act of terrorism (Vilić, 2017). However, there have not been many acts of cyberterrorism (Brenner, 2006).

3.3 Cyberwarfare

Cyberwarfare acts are often seen as political and military motivated war (Cornish et al., 2010). Cyberwarfare uses different weapons in “virtual wars” such as Trojan horses, viruses and worms. These “weapons” are used to disable systems to cause a loss of information or prevent it from functioning. In addition, different software is used to overload Critical Information Infrastructure (CII) to cause disruption or misuse. CII are internet networks or systems responsible for the supervision of CI and the programs that the state and large institutions use (Vilić, 2017).

The common factor between cybercrime, cyberterrorism and cyberwarfare is that the law is often behind and struggles to keep up. Applying existing laws to cyberwarfare is a challenge because of the “weapon” used to carry out cyberwarfare (Robinson, Jones & Janicke, 2015). In addition, there is concern that pursuing cybercriminals across national boundaries can infringe national sovereignty (Goel, 2011). One of the most comprehensive works put together by experts comes from NATO, which describes how existing laws on arm conflict can be applied to cyberwarfare. The Tallinn Manual is not yet lawfully binding, but it can still provide guidance (Robinson, Jones & Janicke, 2015). The following section will discuss past cyberwarfare incidents.

4. Cyberwarfare incidents

This section discusses some of the cyberwarfare incidents that occurred worldwide. Section 4.1 highlights some incidents that occurred in the United States; section 4.2 discusses incidents in the Middle East; section 4.3 highlights incidents in Africa; and section 4.4 highlights incidents in Ukraine.

4.1 Incidents in the U.S.

In 1999 numerous United States government websites were targeted. It was suspected that Chinese hackers were the cause of this attack due to the reported accidental bombing of the Chinese embassy in Belgrade by the U.S. (Goel, 2011).

Another incident that was discovered was in 1998 when U.S. officials discovered an advanced persistent attack (APT) that infiltrated several U.S. CI. The APT known as Moonlight Maze infiltrated the Department of Justice, Department of Defence, and other federal agencies (Kobus, 2016). The infiltration occurred for 2 years, and thousands of unclassified, sensitive information was stolen. This included military hardware designs, military installations, troop configurations and other information on U.S. infrastructure. The Moonlight Maze was traced back to a mainframe computer in Russia, but Russia denied any involvement (Haizler, 2017).

4.2 Incidents in the Middle East

An incident of cyberwarfare that has affected CI was the Stuxnet computer worm that sabotaged Iran’s nuclear facilities (Almeida, Doneda & de Souza Abreu, 2017). This worm was built together by the U.S. and Israel. Stuxnet targets the programmable logic controller (PLCs). PLCs are computers that automate industrial electromechanical processes. Iran discovered the worm in 2010 in their uranium plants. The target of Stuxnet was Industrial Control Systems (ICSs) such as supervisory control and data acquisitions (SCADA) systems. By targeting these CII, Stuxnet gathered information and sent it to an entity that released the information. This caused the final product to be unusable (Kobus, 2016). In addition, the nuclear centrifuges that operated on these ICSs self-destructed due to Stuxnet (Theohary & Rollins, 2015).

In September 2012, a new malware was used in a cyberwarfare act. The incident was caused by a cyberespionage malware called Shamoon. Shamoon targeted companies in the oil and energy sector. One of the companies affected severely was Saudi Aramco, a Saudi Arabian company in the oil and energy sector. Shamoon, a self-replicating and self-injecting malware, attacked 30,000 workstations of Saudi Aramco (Knopová & Knopová, 2014).

4.3 Incidents in Africa

During the Jasmine Revolution in Tunisia, the AMMAR, a government owned Internet Service Provider hacked user accounts for the entire population of Tunisia. In retaliation, adversaries launched Distributed Denial of Services (DDoS) attacks against government and AMMAR websites (Kobus, 2016).

In 2016 two hacktivists’ groups known as #OPAfrica and Anonymous Africa started targeting South Africa due to corruption and perceived social injustices. Anonymous Africa launched DDoS attacks against political parties in Zimbabwe, South Africa and government agencies and companies involved in corruption and the Gupta family. Furthermore, #OPAfrica caused data breaches by attacking South African government systems such as the arms procurement agency, the Government Communications and Information Services, and other vulnerable websites (Van Niekerk, 2018).

4.4 Incidents in Ukraine

Ukraine also experienced a cyberwarfare incident when malware was implanted on Android devices. This malware was used to track and target the Ukrainian artillery units. The tracking occurred from late 2014 through 2016. It retrieved locational data and communications data from these artillery units. The hacking group

responsible for this was linked to the Russian government. It was reported that 9000 users had the malware running on their applications (Gazula, 2017).

Another incident occurred again in Ukraine in December 2015. Blackenergy, a Trojan, infected the Ukrainian electricity distribution network. This attack interrupted more than 225,000 people's electricity supply. Furthermore, it was reported that Blackenergy was intended to be used during the electoral period in Ukraine to destroy videos and image files that would influence the election (Izycki & Vianna, 2021). These incidents of cyberwarfare have caused adverse effects on the CI it has targeted. Table 1 summarises these cyberwarfare incidents and evaluates why these incidents can be classified as cyberwarfare. The following section discusses the possible effects of cyberwarfare on CI.

Table 1: Summary of cyberwarfare incidents and why they can be classified as such

Incident	Reason this is a Cyberwarfare incident
Attack on U.S. government websites by Chinese hackers	The target of the attack was a government website caused by a non-state actor. The attack was due to political reasons, and the attack caused severe disruptions.
The Moonlight Maze APT attacked several U.S. CI.	The target of the attack was a state's CI, and the attack was believed to be caused by another state (Russia). The attackers stole sensitive information that caused political issues.
The Stuxnet worm sabotaged Iran's nuclear facilities.	This attack is cyberwarfare because the target of the attack was a state's CI: Iran's nuclear facilities. The destruction of the nuclear centrifuges indicates a cyberwarfare incident since cyberwarfare incidents typically aim to cause disruption or destruction to CI.
Shamoon malware targeted companies in the oil and energy sector, including Saudi Aramco.	The target of the incident was the oil and energy sector, which qualifies as CI for many nations. The disruption of the CI indicates that this was a cyberwarfare incident.
DDoS attacks were launched against the government and AMMAR, a government-owned Internet Service Provider in Tunisia.	This is a cyberwarfare incident because the target of the DDoS attacks were the Tunisian government websites and a state-owned company. The attacks, which caused serious disruptions, were due to political reasons. The attacker was a non-state actor.
Anonymous, a Hacktivists group, launched DDoS attacks against Zimbabwean and South African government agencies, political parties, and companies. #OPAFrica caused data breaches on South African government websites.	The attacks qualify as cyberwarfare because they specifically targeted government websites. While the attacker was not a state actor, the attacks still caused disruptions on the nations' systems, which is often the aim of cyberwarfare.
A malware was implanted on Android devices that tracked and targeted Ukrainian artillery units	This is a cyberwarfare incident because Russia, a state actor, was linked to this attack against another state, Ukraine. This attack retrieved information from Ukraine's CI by tracking and targeting the Ukrainian artillery.
Blackenergy infected the Ukrainian electricity distribution network.	Cyberwarfare usually targets CI. This incident targeted the electricity distribution network, Ukraine's CI. The incident caused a significant disruption that affected more the 225,000 people.

5. Techniques to protect Critical Infrastructure from Cyberwarfare

The following section discusses some techniques that can be used to protect CI from cyberwarfare. Section 5.1 discusses the system modelling technique, section 5.2 discusses the defence-in-depth strategy, and section 5.3 highlights techniques related to human aspects of security.

5.1 System Modelling

Due to the high complexity of CI, it is necessary to model the complexity of these systems. Modelling can help study and identify vulnerabilities in the CI. This can be done through designing virtual environments which can provide a simulation of features of CI (Merabti, Kennedy & Hurst, 2011). In addition, attacks can also be simulated through war games. War games include simulating attacks on the SCADA system, which helps mitigate attacks because it helps identify weaknesses (Nicholson et al., 2012). The United States launched a war game in 2010 February. This war game was known as Cyber Shockwave (Nelson, 2011). The war game infected 5 million smartphones using a Trojan. This infection then manifested into a DDoS attack that affected government

systems. The results of the war game simulation revealed that the U.S. was not prepared to defend itself against a cyberwarfare event (Nicholson et al., 2012). Therefore, such simulations (as tools used in system modelling) can improve the protection of CI.

5.2 Defence in depth

CI should adopt a defence-in-depth strategy. This requires having multiple layers of security such that if an attack penetrates one layer, it cannot penetrate the next layer. These layers employ different technologies and intrusion detection systems (IDS) to secure CI (Hurst, Merabti & Fergus, 2014). The technologies that can be applied can be general preventative measures used in general IT systems such as anti-virus, firewalls, IDS etc. (Nicholson et al., 2012). The defence in-depth needs to implement three levels of security: low, medium, and high. Medium and high levels of security are designed specifically for employees such as managers and system administrators who has access to sensitive information systems and infrastructure assets. In contrast, the low levels of security are for employees who only have basic access (Hurst et al., 2014).

CI infrastructure that adopts this defence-in-depth strategy can also employ Unified threat management (UTM) systems (Hurst et al., 2014). UTM systems provide protection for network, hardware, and software layers. It achieves this by combining several security technologies such as pattern recognition systems, firewalls, IDS and analysis middleware (Zhang, Deng, Chen, Xue, & Lin, 2010). UTM systems can secure CI because they are a unified architecture with different security techniques and are easy to deploy (Hurst et al., 2014).

5.3 Human aspects

Humans are often the weakest link in the security of systems. Security policies, procedures, and protocols need to be defined to prevent attackers from exploiting humans (Nicholson et al., 2012). Training and awareness programs on security also need to be applied.

In addition, there is often a shortage of security experts that can secure CI from cyberwarfare. It is in the best interests of governments to invest in specialised security training to create a talented workforce that can help a nation protect against cyberwarfare from other nations (Tiirmaa-Klaar, 2016). Therefore, training and education to create security experts can protect CI against cyberwarfare.

6. Possible effects of Cyberwarfare on Critical Infrastructure

CI is often controlled by remote systems known as Industrial Control Systems (ICSs). ICS also include SCADA systems (Mcginthy & Michaels, 2019). Cyberwarfare targets ICSs to threaten the integrity, confidentiality, and availability of these systems. If an attack successfully achieves this, it can have dangerous effects such as disruption, alteration, or destruction of CI (Montgomery, 2018).

Cyberwarfare can disrupt CI by attacking the integrity of the SCADA systems. For example, if a cyberwarfare event targets and changes the control signals of the SCADA systems, it can cause it to malfunction, which can also affect the availability of the system (Kovacevic & Nikolic, 2015). If a SCADA system becomes unavailable, the CI it manages (i.e., power distribution, water and waste distribution services) will be disrupted. If these services are disrupted, it can result in major economic losses and affect the livelihood of citizens (Thakur, Ali, Jiang & Qiu, 2016). The disruption caused by attacking SCADA systems was seen in the 2010 attack of the Stuxnet worm on Iran's uranium plants discussed in section 4.2 (Kobus, 2016). If a cyberwarfare incident disrupts a single CI sector, the effects of this attack can propagate to other CI sectors. This can occur due to the interdependencies and dependencies among the different CI sectors (Kovacevic & Nikolic, 2015). For example, water management systems require electric power generation infrastructure to supply water. Therefore, cyberwarfare attacks can spread to different sectors, which is a significant threat to CI.

The disruption that cyberwarfare can cause can persist for long durations of time because of the difficulty to perform updates and software patches on SCADA systems (Cárdenas, Amin, Lin, Huang, Huang, & Sastry, 2011). Before patches are applied, they must be thoroughly tested, and performing upgrades requires planning and placing the systems offline (Kovacevic & Nikolic, 2015). Therefore, the effects of cyberwarfare can last for a long time.

Cyberwarfare attacks can alter CI by breaching the confidentiality and integrity of data supplied to the central monitoring systems that manage CI. A cyberwarfare attack could supply inaccurate data to these systems through a virus. This could result in the SCADA system reacting to this false data, which can cause the CI to

perform dangerous actions (Nicholson, Webber, Dyer, Patel & Janicke, 2012). In addition, cyberwarfare can destroy CI by causing failures to the SCADA systems, and that can result in causing damage to physical systems that can be irreparable. It can lead to consequences on the public safety and health of citizens (Kovacevic & Nikolic, 2015). Therefore, the effects of cyberwarfare on CI can affect the physical world, and one of the major concerns that the failure of CI could cause is the loss of human life (Nicholson et al., 2012). Due to the number of dangerous effects that cyberwarfare can cause, it is essential to apply techniques to protect this CI. But with the threats of cyberwarfare and protection techniques, is South Africa ready to protect itself against a cyberwarfare incident? The following section discusses this.

7. South Africa's readiness for Cyberwarfare

South Africa (SA) is not prepared to defend itself against a cyberwarfare attack. While there has not been a major reported cyberwarfare incident, the number of cyberattacks in South Africa have increased due to increased connectivity (Peter, 2017). SA's responses to these attacks demonstrated that it is not prepared. For example, the recent attack on Transnet's information technology networks severely disrupted Transnet's CI services. Transnet is an SA-owned rail, port and pipeline company that experienced disruptions that interrupted their cargo movement. As a result, cargo terminals at Cape Town stopped operating, resulting in Transnet switching to manual processes. Transnet's response to switching to a manual approach indicates that Transnet was not prepared for an attack targeting their CI. Unfortunately, public information is limited, making it difficult to determine what security technique to protect their ports (Reva, 2021).

Another incident was the attack on the City of Johannesburg (CoJ) systems. CoJ experienced a network breach that allowed unauthorised access to their information systems. As a result of the attack, CoJ had to shut down customer-facing systems such as e-services (Moyo, 2019). CoJ shutting down a customer-facing system as a response showed unreadiness for an attack of that nature. In this case, applying defence-in-depth could assist CoJ in protecting their CIs since having multiple layers of security could prevent the network breach from penetrating other layers of the system.

Another attack that further highlights SA's unpreparedness against cyberwarfare is the ransomware attack that targeted 66 Life Healthcare hospitals. The ransomware attack caused several administrative delays forcing the hospitals to switch to manual processing systems (Bottomley, 2020). Such an attack could be mitigated by applying system modelling (to identify vulnerabilities), defence-in-depth (to add multiple layers of security) and human aspects (such as staff training and awareness programs).

In addition, SA's lack of cyberwarfare readiness is due to South Africa's legislative systems. In addition, there is currently a lack of international cyber defence treaties that specifically address cyberwarfare (Grobler & van Vuuren, 2012). The lack of treaties makes it extremely difficult to prosecute criminals from different states (Goel, 2011) and contributes to South Africa's lack of readiness because there is no legal agreement on the consequences that attackers who commit cyberwarfare can face. While South Africa has taken legal steps to protect against cyberattacks by passing legislation such as the Cyber Crimes Act (Moyo, 2021), it is still not enough to prepare South Africa for cyberwarfare. The Cyber Crimes Act covers cybercrime and not cyberwarfare, which are two different cyberattacks performed for different reasons, as discussed in section 3.

Qualified professionals are leaving South Africa for opportunities overseas, which means South Africa is losing a valuable workforce in sectors such as infrastructure and development (Thompson, 2021). From section 5.3, it was noted that skilled experts in fields such as security are needed to protect against cyberwarfare. But since South Africa is losing experienced professionals, which puts further strain on effectively preparing for a cyberwarfare incident. Skilled staff is needed to implement technologies to create resilient CII. Since there is a lack of an experienced workforce, technologies for protecting CII have a higher likelihood of being outdated. Therefore, South Africa might not have the right tools and approach against a cyberwarfare incident. The following section will conclude the paper.

8. Conclusion

Different concepts were discussed to meet the objective of the paper. The main aspects addressed were first understanding cyberwarfare by considering various definitions. Cyberwarfare is seen as actions whose aim is to bypass a nation's networks or computers that can result in significant damage or disruption. There are currently different views of cyberwarfare. Some definitions describe that cyberwarfare actions are only between nation-

state actors (Clarke & Knake, 2010). Other definitions specify that it can also involve non-state actors (Robinson, Jones & Janicke, 2015). Cyberwarfare can be confused with other cyberattacks such as cybercrime and cyberterrorism. Therefore, the paper discussed each of the three concepts. It was found that cybercrime is often committed for personal reasons. In contrast, cyberterrorism is committed due to political reasons to cause fear, similar to a physical terrorist attack.

Several incidents were discussed to understand the dangers of cyberwarfare, including why they were classified as cyberwarfare. Incidents such as Stuxnet highlighted some of the possible effects of cyberwarfare on CI. Due to these effects, techniques that can protect CI are required. Techniques like system modelling can help identify weaknesses in CI (Nicholson et al., 2012). Moreover, human aspects are essential since humans are usually the weakest link in cyberattacks. Effects such as disruption and damage to a nation's CI were also discussed to show that cyberwarfare can have a devastating impact on CI. Lastly, the paper looked at whether South Africa is prepared to defend itself against cyberwarfare. Our opinion that South Africa is not ready was supported by how South Africa responded to cyberattacks incidents such as the Transnet attack. In addition, the lack of legal treaties addressing cyberwarfare contributes to South Africa's lack of readiness. Therefore, the objective of the paper, which was to discuss cyberwarfare and its effects on CI, was met.

References

- Almeida, V.A.F., Doneda, D. and de Souza Abreu, J. (2017) "Cyberwarfare and digital governance", *IEEE Internet Computing*, Vol. 21, No. 2, March, pp 68-71.
- Bernik, I. (2014) *Cybercrime and cyber warfare*, John Wiley & Sons, Hoboken.
- Bottomley, E. J. (2020) "SA hit as hackers target hospitals during Covid-19 crisis- here's what Life may be facing", [online], Business Insider, [SA hit as hackers target hospitals during Covid-19 crisis – here's what Life may be facing \(businessinsider.co.za\)](https://www.businessinsider.co.za/sa-hit-as-hackers-target-hospitals-during-covid-19-crisis-here-s-what-life-may-be-facing/).
- Brenner, S.W. (2006) "Cybercrime, cyberterrorism and cyberwarfare", *Revue internationale de droit Zenal*, Vol. 77, No. 3, pp 453-471.
- Broadhurst, R., Grabosky, P., Alazab, M., Bouhours, B. and Chon, S. (2014) "An analysis of the nature of groups engaged in cyber crime", *International Journal of Cyber Criminology*, Vol. 8, No. 1, January-June, pp 1-20.
- Cárdenas, A.A., Amin, S., Lin, Z.S., Huang, Y.L., Huang, C.Y. and Sastry, S. (2011) "Attacks against process control systems: risk assessment, detection, and response". *Proceedings of the 6th ACM symposium on information, computer and communications security*, March, pp 355-366.
- Chandra, A. and Snowe, M.J. (2020) "A taxonomy of cybercrime: Theory and design", *International Journal of Accounting Information Systems*, Vol. 38, September, pp 1-20.
- Clarke, R.A. and Robert, K.K. (2010) *Cyber war: the next threat to national security and what to do about it?*, Harper Collins.
- Colarik, A.M and Janczewski, L.J. (2012) "Establishing cyber warfare doctrine", *Journal of Strategic Security*, Spring Vol. 5, No. 1, pp 31-48.
- Cornish, P., Livingstone, D., Clemente, D. and Yorke, C. (2010) *On cyber warfare*, Chatham House, London.
- Gazula, M.B. (2017). *Cyber warfare conflict analysis and case studies*, Massachusetts Institute of Technology, Cambridge.
- Goel, S. (2011) "Cyberwarfare: connecting the dots in cyber intelligence". *Communications of the ACM*, Vol. 54, No. 8, August, pp 132-140.
- Grobler, M. and van Vuuren, J. J. (2012) "Collaboration as proactive measure against cyber warfare in South Africa", *African Security Review*, Vol. 21, No. 2, June, pp 61-73.
- Haizler, O. (2017) "The United States' cyber warfare history: Implications on modern cyber operational structures and policymaking", *Cyber, Intelligence, and Security*, Vol. 1, No. 1, January, pp 31-45.
- Hurst, W., Merabti, M. and Fergus, P. (2014) "A survey of critical infrastructure security". *International Conference on Critical Infrastructure Protection*, Vol. 441, pp 127-138, Springer, Berlin, Heidelberg.
- Izycki, E. and Vianna, E.W. (2021) "Critical Infrastructure: A Battlefield for Cyber Warfare?", *ICCWS 2021 16th International Conference on Cyber Warfare and Security*, pp 454-464.
- Knopová, M. and Knopová, E. (2014) "The Third World War? In The Cyberspace. Cyber Warfare in the Middle East". *Acta Informatica Pragensia*, Vol. 3, No. 1, pp 23-32.
- Kobus, J. (2016) *Cyberwarfare: The evolution of war*, Utica College, Utica.
- Kovacevic, A. and Nikolic, D. (2015) "Cyber attacks on critical infrastructure: Review and challenges", *Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance*, pp 1-16, University of Belgrade, Serbia.
- Mcginthy, J. M. and Michaels, A. J. (2019) "Secure industrial Internet of Things critical infrastructure node design", *IEEE Internet of Things Journal*, Vol. 6, No. 5, pp 8021-8037.
- McGuire, M. and Dowling, S. (2013) *Cybercrime: A review of the evidence: Summary of key findings and implications*, Home Office Research Report 75, London: Home Office.
- Merabti, M., Kennedy, M., & Hurst, W. (2011) "Critical infrastructure protection: A 21 st century challenge", *2011 International Conference on Communications and Information Technology (ICCIT)*, pp 1-6.
- Montgomery, M. (2018) "Proliferation of cyberwarfare under international law: virtual attacks with concrete consequences", *S. Cal. Interdisc. L.J.*, Vol. 28, pp 499-521.

- Moyo, A. (2021) "President Ramaphosa signs Cyber crimes Bill into Law", [online], I.T. Web, <https://www.itweb.co.za/content/LPp6VMrDJJovDKQz>.
- Moyo, A. (2019) "City of Johannesburg hit by cyber attack", [online], ITWeb, <https://www.itweb.co.za/content/dgp45qaG8gZ7X9l8>.
- Mueller, M. L. (2020) "Against sovereignty in cyberspace", *International Studies Review*, Vol. 22, No. 4, 20 September, pp 779-801.
- Nelson, C. (2011) "Cyber Warfare: The Newest Battlefield", [online], Computer Science & Engineering Washington University in St.Louis, <https://www.cse.wustl.edu/~jain/cse571-11/ftp/cyberwar/>.
- Nicholson, A., Webber, S., Dyer, S., Patel, T. and Janicke, H. (2012) "SCADA security in the light of Cyber-Warfare", *Computers & Security*, Vol. 31, No. 4, June, pp 418-436.
- Parks, R. C. and Duggan, D. P. (2011) "Principles of cyberwarfare", *IEEE Security & Privacy*, Vol. 9, No. 5, 26 September, pp 30-35.
- Peter, A. S. (2017) "Cyber resilience preparedness of Africa's top-12 emerging economies", *International Journal of Critical Infrastructure Protection*, Vol. 17, March, pp 49-59.
- Reva, D. (2021) "Cyber attacks expose the vulnerability of South Africa's ports", [online], Institute For Security Studies, <https://issafrica.org/iss-today/cyber-attacks-expose-the-vulnerability-of-south-africas-ports>.
- Robinson, M., Jones, K. and Janicke, H. (2015) "Cyber warfare: Issues and challenges", *Computers & security*, Vol. 49, March, pp 70-94.
- Sarre, R., Lau, L. Y. C. and Chang, L. Y. (2018) "Responding to cybercrime: current trends", *Police Practice and Research*, Vol. 19, No. 6, 20 September, pp 515-518.
- Thakur, K., Ali, M. L., Jiang, N. and Qiu, M. (2016) "Impact of cyber-attacks on critical infrastructure", *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity)*, *IEEE International Conference on High Performance and Smart Computing (HPSC)*, and *IEEE International Conference on Intelligent Data and Security (IDS)*, pp 183-186.
- Theohary, C. A. and Rollins, J. W. (2015) *Cyberwarfare and cyberterrorism: In brief*, Congressional Research Service, Washington.
- Thompson, W. (2021) "Brain drain a threat to economic recovery, warns RMB boss", [online], BusinessDay, <https://www.businesslive.co.za/bd/economy/2021-02-10-brain-drain-a-threat-to-economic-recovery-warns-rmb-boss/>.
- Tiirmaa-Klaar, H. (2016) "Building national cyber resilience and protecting critical information infrastructure", *Journal of Cyber Policy*, Vol. 1, No. 1, March, pp 94-106.
- Van Niekerk, B. (2018) "Information warfare as a continuation of politics: An analysis of cyber incidents", *2018 Conference on Information Communications Technology and Society (ICTAS)*, pp. 1-6.
- Vilić, V. M. (2017) "Dark web, cyber terrorism and cyber warfare: Dark side of the cyberspace", *Balkan Social Science Review*, Vol. 10, No. 10, October, pp 7-25.
- Zhang, Y., Deng, F., Chen, Z., Xue, Y. and Lin, C. (2010) "UTM-CM: A practical control mechanism solution for UTM system", *2010 International Conference on Communications and Mobile Computing*, Vol. 1, pp 86-90.

Addressing the Skills Shortage in Cybersecurity

Gareth Davies, Angela Mison and Peter Eden

University of South Wales, Treforest, UK

Gareth.Davies@southwales.ac.uk

Angela.Mison@southwales.ac.uk

Peter.Eden@southwales.ac.uk

Abstract: Security by design is an up and coming paradigm which seeks to decrease the opportunity for corruption and disruption, and to increase the inherent stability, dependability and resilience of systems. Cyber experts need to be involved in the design phase as Finagle's Law, interpreted for cybersecurity, is 'if it can be hacked, it will – at the worst possible time'. Testing is designed to counter Murphy's Law and reduce resistentialism. Defensive programming, original cybersecurity, carries an overhead, for which there is often no demonstrable return. The priority of the design stage cybersecurity expert is to plan for contingencies and think like a hacker. It is about risk management; understanding this dictates security by design requirements in the knowledge that interconnected systems' security is only as strong as its weakest link. Definition and audit of Service Level Agreements (SLA) are an essential part of cybersecurity, as is the audit of any third party suppliers of componentry of the system under design. Governance and policy definitions, exciting for some, are integral to cybersecurity. Beyond this, the cybersecurity expert must consider system failure. Recent ransomware attacks have demonstrated the necessity for business continuity plans as recovery has still taken time. The Blue Team Field Manual has an impressive list of necessary documentation and actions required in this event, but glosses over the effort required. While the above is rolling maintenance, threat hunting differs. Every vulnerability or threat must be evaluated for consequential impact. Either a passionate interest in psychology or an extremely jaundiced view of the world is a necessary attribute for cybersecurity. Reality is so different from the aspirations of potential pen testers, incident responders, and AI security engineers facing 3 – 5 post-graduation years to proficiency, it is no wonder disillusion results in a shortage of 3.5 million cybersecurity experts.

Keywords: Murphy's Law, Finagle's Law, Resistentialism, Cybersecurity, Career path

1. Introduction

Everyone, everywhere, will have come across Murphy's and Finagle's Laws, and experienced resistentialism. To understand how they relate to cybersecurity requires their definition.

Murphy's Law – *If it can go wrong it will*, although the original, stemming from 1948, is stated as *if there is more than one way to do a job, and one of those ways will result in catastrophe, then somebody will do it that way* (Murphy, 2012). It remains a tenet of defensive design, and translated to the cybersecurity context - *if it can be hacked it will*.

Finagle's Law of Dynamic Negatives – *if it can go wrong it will – at the worst possible time* (techopedia, 2021), is a corollary to Murphy's Law giving recognition to the inevitability of negative outcomes occurring, but which, due to timing, amplify adverse effects.

Resistentialism – *the seemingly spiteful behaviour manifested by inanimate objects where those objects cause problems and exhibit malice to humans* (Jenning, 1948)

Cybersecurity – and now comes the difficulty. There are many definitions, most of which emphasise the technical:

Cybersecurity refers to the body of technologies, processes, and practices designed to protect networks, devices, programs, and data from attack, damage, or unauthorized access. Cybersecurity may also be referred to as information technology security. (Digital Guardian, 2021)

It all seems do-able, until you get the definition from people who really know what they are talking about:

Cybersecurity is how individuals and organisations reduce the risk of cyber attack. (National Cybersecurity Centre, 2021)

Cybersecurity has suddenly gone from devices, networks and all things technical, to something amorphous, waffly, and boring. It permeates the entire organisation and culture, where the buck stops with the Board who,

in the UK, are responsible and accountable, potentially personally and financially liable due to inadequate reporting of solvency and risk (Financial Reporting Council, 2016), both of which can arise from cybersecurity failings. The least that can happen in response to a publicly advertised security breach is loss of reputation for the organisation concerned. Where quoted, the market demonstrates loss of confidence, with an immediate reduction in share price,

Although addressing the banking industry, The Financial Conduct Authority (Steward, 2021) Discusses the role of self interest of senior managers and their potential personal liability. Seen from this perspective cybersecurity is not about securing anything, it is about risk management, both personal and organisational, and organisational continuity. When an organisation suffers a cyber-attack, one has to ask why cybersecurity is held responsible for a consequential organisational failure in addition to the technical failure. If a house falls down, is it the electrician's fault? When the skills shortage is discussed, it is not certain exactly what skill is in short supply and whether some skills could and should be available from non-technical areas.

It has been shown that risk management by itself cannot provide a complete solution yet low-level technical graduates are being asked to make decisions about risk mitigation and business continuity, often when cybersecurity has failed. Those graduates thought their jobs would be exciting, like gaming, fighting black hats and fixing battle damage. There is a mismatch of expectation, responsibility, and accountability.

Technical cybersecurity is a post hoc system development add on - on par with extended testing, proving the presence of errors, never their absence. It is necessitated by the inevitable inadequacy of the system design phase, failure of the resultant implementation, lack of business planning and lack of top-management understanding of the consequences of business decisions in an interconnected world. Incident response is the reaction to failure of cybersecurity. In reality, it can lead to organisational upheaval.

2. The real purpose of cybersecurity

Cybersecurity addresses, in part, technical systems. Its broader definition (*ibid*) comes from viewing systems in their context. Only truly failed systems exist in a vacuum - implemented, but not installed. Systems are developed to support something. In the manner of an inverted triangle, what they support is some type of purposeful organisation, thus cybersecurity's goal is to enable continuity of the organisation dependent on that system.

The analogy of the inverted triangle is useful to define purposes and consequences. An upstanding inverted triangle is inherently unstable. The organisation can be represented by the whole while the technical systems are at the base point. Tamper with the base, and the whole or part may topple, taking with it continuity of operation. The real purpose of cybersecurity is to prevent that toppling and ensure continuity if the triangle does topple. Incident response is equivalent to FEMA after the earthquake, concerned with clean up and rebuilding, with prevention of repetition in mind.

3. Cybersecurity in the organisation

3.1 Specialist selection – team portfolios, diversity, and human resources

There is a requirement for detailed definitions of roles within an organisation. These normally encompass the necessary skill sets (NICCS, 2021). Forward looking organisations recognise that any successful team requires a portfolio of skills and mind sets (Patterson, 2018). When examining a vacancy, it is necessary to identify both the skills that are missing, the mind set which will enhance team performance, and then to describe that unicorn. A passionate interest in psychology or an extremely jaundiced view of the world may be a necessary attribute together with disillusion caused by Murphy's law.

Formulaic selection can miss the greatest talents and closest matches to an organisation's actual needs which are often based on less obvious factors and which, unexpectedly, may be found in-house in a different field. Unfortunately, outsourcing and automated selection techniques can be the same as asking the wrong question, where the answer does not matter, and consequently, a limited talent pool is presented. Currently, it is reported that 99% of Fortune 500 companies use some form of applicant tracking systems (Qu, 2019)

Cybersecurity resourcing supports the argument for a Chief Neurodiversity Officer (Levell, 2017). It maybe that the inherent mind set of an applicant is more important than skills which can be learnt, extended or covered by

others, and more deterministic of suitability of a candidate and the ultimate career path. Is a detail oriented person required, or is it someone whose mind is buzzing with half-baked ideas? A good recruiter knows who, in his team, has these traits, and whether more of the same, or different traits are needed.

3.2 Security by design – architects, designers and risk managers

Security by design seeks to reduce the opportunity for corruption/disruption of a system and to increase its inherent stability, dependability and resilience. There is a need for the involvement of cyber experts in the design phase. Murphy's Law of computing systems failing at introduction should resonate. Given a choice of options in a requirements specification, the developer is not necessarily the best person to determine the direction of the organisation through an apparently logical, but ultimately disastrous choice. It must be pointed out that cybersecurity experts are exempt from neither Murphy's nor Finagle's Law.

The priority of the developer is getting the system to fulfil the requirements. Defensive programming, the well understood original cybersecurity, can carry a large overhead for which there is often no demonstrable return, and which, if developed without consideration of psychological consequences, can increase any inherent resistentialism. Good user interfaces result from extensive research..

The role of the cybersecurity expert at this stage is to identify what the system can be made to do, define contingencies, the 'what if', scenarios, being aware of the threat landscape and thinking like a hacker. It behoves the cybersecurity expert to believe the worst of the world and to dictate the requirements for extended defensive programming in order to defend against the foreseeable.

Worse than this, the cybersecurity expert needs to understand the context in which the system operates - an increasingly interconnected world with complex supply chains. The security of an interconnected system is only as strong as its weakest link, and organisational hubris may lead to the assumption that this is elsewhere. An internal weakness may expose the organisation to, at least, compensation claims from interconnected parties, while a suspected weakness in an interconnected party may suggest intrusive investigation.

Risk management in system development requires an unusual skill set dealing with evaluation of consequences and mitigation decisions. The cybersecurity version of this includes the answer to 'if this failed, what would we need to do to keep operating?'. That answer incorporates continuity which should be added to any process documentation – not obviously a cybersecurity role, but in the broad definition, yes, it is.

Cybersecurity adds to cost and is subject to criticism. The difficulty is one faced by intelligence Services: how to appreciate and value the effort of protecting against something that does not happen, but which event is a possibility? If the event did happen, it could be organisationally existential.

3.3 Murphy's and Finagle's Laws and resistentialism – testers

It is a sad fact that development projects are constrained. There are only three factors to consider in a project: time, resources, and money. Some would argue quality, but that is taken as a given here. Where a project is running late, testing is usually the economic sacrifice. Testing proves the presence of errors, never their absence. In pure system testing, it is the place to find errors and iron out the bugs and irritations attributable to Murphy's Law and resistentialism.

An additional, separate test suite is required to meet the requirements of security by design. Whether this is a formalised test suite based on scenarios, and/or letting the technical experts try a little pen testing is cultural. The testing must take place in context, i.e. its interconnected state, and this would be the ideal time for ethical pen testing of partners.

3.4 Supply chain and interconnected parties – procurement, quality, legal, risk, auditors

The position is simple; if your organisation can hack into interconnected systems, so can someone else, which opens your organisation to potentially unconsidered risk. Of course, having the ability to do this depends on your position in the food chain, the leverage you can exert, and whether your ethics, contracts, or Service Level Agreements (SLA) with others permit arrogation of their systems.

Discussion of SLA with interconnected parties has not been seen as the preserve of the cybersecurity expert. Audit of compliance with the terms of the SLA is an essential part of cybersecurity, as it will be structured to ensure those terms. Beware: compliance does not equate to performance!

Resilience, dependability and integrity requirements may require the continuous monitoring or audit of any third party suppliers of componentry or systems under consideration. Where there is a less than cooperative party, there is a higher risk and arrogation may be necessary to ensure cybersecurity and potential continuity of operation. In a third party supplier relationship, the ongoing nature of their business, continuing longevity, quality and rights to fundamental materials and documentation in case of their cessation of trading or existence may be elements for inclusion in contracts or SLA. This inclusion protects the maintenance and continuing operation of said systems in the short term.

The skill sets for definition and audit of such elements may not be the preserve of the cybersecurity expert, but any organisation of note will have a legal, financial, procurement departments capable of cooperating to include the requisite cybersecurity requirements. Collaborative working of this nature increases the cybersecurity expertise and culture in these disparate areas and increases the organisational awareness of the cybersecurity expert. This exchange and extension of knowledge is beneficial both personally and organisationally.

3.5 Policy and Governance – lawyers, psychologists, human resources, change managers

Policy and governance changes are often the result of an event and designed to prevent repetition. For most, policy and governance definitions are not exciting parts of the cybersecurity remit, but remain essential. To technical people, change management has hitherto been associated with endless documentation arising from the change to one or more lines of coding. Furthermore, the cybersecurity expert must consider system failure, at which point, all hell breaks loose, but which hell must somehow be constrained into processes and procedures, and communicated within the organisation.

The banking industry is addressing conduct and culture. Interestingly, (Group of 30, 2018) states:

Conduct is not just about purposeful misbehaviour, but also unintended consequences from decisions and/or lack of skills and knowledge.

Policy and governance actually dictate the way personnel should behave and the way the enforcement regime. Understood slightly differently, this equates to the culture of the organisation. From this perspective, any change, to a policy or its governance, represents a cultural or behavioural change, usually not attainable by diktat, but more by persuasion and education. Habitually there is resistance to change, thus change management has a skill set in its own right, addressing the needs of personnel and smoothing transitions wherever possible.

An example of a novel cybersecurity policy which may encounter resistance is the termination of use of personal devices or personal internet usage. There are very good cybersecurity reasons for these particular policies, but given the attachment people have to their devices, or web surfing, they are policies which may meet with resistance and need careful management, education and reinforcement to achieve a more cyber secure workplace. This example demonstrates that edicts, emanating from the basement, where the cybersecurity 'geeks' hang out and are detached from reality, can have profound effects on organisational behaviour, culture and even induce organisational churn. Having said that, it highlights the 'us and them' mentality that may exist.

This is a relatively obvious example. It should be borne in mind that any policy and/or governance changes affect everyone in the organisation and thus the important consideration is how to achieve their adoption without disruption. The interaction of change managers and cybersecurity experts may be beneficial, increasing the empathetic skills of the cybersecurity personnel, and leading to informal communication with relevant personnel preceding the release of a new or changed policy – sometimes wording and tone matter, as does an understanding that the language used must be appropriate for its audience.

3.6 NOCs, SOCs – technical experts

Attacks do not need to be direct, they can be consequential, from an attack on a third party attempting to corrupt or manipulate something entirely different, somewhere else. Going back to the system design and development phase, what can the system be made to do, and if you were a hacker, how might that be

monetised? Every identified vulnerability or threat must be evaluated in this light. As with all things cyber, risk management cannot stop at the borders of the organisation.

The NOCs and SOCs are similar to continuous testing and monitoring for those Murphy's and Finagle's Laws events, the one's for which there is no contingency. There is an unfounded belief in project and programme management that where there is a contingency plan for an event, that event will not happen, but some other event will. This adequately describes the behaviour of IDS and IPS. Some of these events may be induced through internal system errors or introduced by malefactors or inattentive interconnected partners so the cybersecurity personnel have to adapt to detecting and exercising their expertise in a high pressure environment, recognising that no matter what they do, they will be held to blame for any organisational failure. They need to be protected from unrealistic demands and harrying regarding any progress and this can be provided by the cybersecurity incident response team (CSIRT) processes and procedures allied with continuity plans.

4. Incident Response – the organisation

4.1 Incident Response – technical experts

Cybersecurity experts are not exempt from Murphy's Law. It is always assumed that cybersecurity will fail, otherwise there would be no pre-ordained CSIRT or procedures. While something is floating around the network, it cannot do too much damage, but if it becomes active in any node, there is a problem. Incident response is the answer to the failure of cybersecurity.

Often the malefactor's intent does not manifest to defenders until there is an observable event which is detected in the system or an organisational functional area by means of unexpected behaviour of the system. Periods for detection can be protracted due to user expectation of resistentialism, or belief that an update has taken place which has altered the behaviour of the system. 'The system is slow today' can hide a multitude of problems. Reporting is problematic.

Bearing in mind the speed of the internet data transmission is of the order of 178 terabits per second (Nield, 2020), and a round trip between Europe and the USA takes approximately 70 milliseconds (Brown, 2013), it is useful to consider a domino race analogy, applied to the topology, understand where potential break points might be defined and the near impossibility of stopping anything in the time available once it is past the IDS. It is impossible to fix anything unless you have an idea of what it is, i.e. what it might be doing and where it might be. Anyone with experience of mission critical systems will recognise that any bug fix can only take place when the bug has been found, no matter what the SLA says.

4.2 Incident Response – technical experts, lawyers and public relations

The incident response functions are activated on the recognised failure of the cybersecurity system and process which has a consequential impact on continuity. Post event, it takes on a forensic aspect, followed by a clean-up and revisions to policies and procedures. It can never be assumed that all traces of an event have been eradicated.

If the organisation is outward facing or there are interconnected parties, legal and communication issues arise. There are a number of public facing organisations who have dealt with this over the years, and the continuing difficulty is in the determination of timing of normal service resumption, the parties who should be notified, any lasting damage to systems, and the effect on continuity. There is a current assumption that any public notification of website technical issues or reduction of system functionality due to updates are the result of a hack. Whether this is true, it represents a major inconvenience to any users and may be detrimental to the organisation.

4.3 Business continuity – the Board, finance, legal, and essential organisation functions

Recently, the necessity for business continuity plans has been demonstrated as a result of ransomware, but it has still taken time to overcome disruptions, and to recover. The Blue Team Field Manual (White & Clark, 2017) has an impressive list of necessary documentation and actions required in the event of cybersecurity failure. It does not indicate the effort required to complete the documentation, but the knowledge and expertise involved in their production indicates that it is extensive. For an organisation to produce such documentation is a major project.

The continuity plan that is contained within the documentation may deal with the complete failure of organisational systems eg ransomware, when continuity may mean reverting to manual processing. Off-site, disconnected back-up systems may be instantiated but, depending on back-up frequency, there may be a loss of data in the interim. There may be plans to isolate the organisation from authorised interconnected third parties when the necessary continuity plan may mean manual operation of parts of the business, for example supply chain, or alternative bulk data exchange mechanisms. How the continuity plan is implemented, depends on the definition of the mission critical parts of the organisation function and non-corruption of associated critical network paths.

Continuity management (CM), an organisational responsibility, implies that continuity plans are in place. When cybersecurity fails, the situation frequently degenerates to crisis management, not at all the same thing. There is a marked difference in the results achievable when an ordered rather than chaotic approach is taken as noted by the Civil Contingencies Act Enhancement Programme (Cabinet Office, 2012). At no time should organisational continuity be the sole responsibility of cybersecurity.

4.4 Maintenance – the organisation

Maintenance or changes, recommended or mandatory, are usually the result of some event and incident response pursuant to a post event audit, similar to a forensic analysis of processes and procedures. Updating the relevant systems and/or paperwork is undertaken in an attempt to prevent a repetition of that event which caused a cybersecurity failure and to improve the response to such a failure. Its primary purpose should never be apportioning blame.

Maintaining documentation relating to policies, governance, contracts, SLA, processes, procedures, key staffing lists etc can be a much hated task to the cybersecurity expert. Any chore which is much hated is unlikely to be done well, so revelling in such a detail oriented task is a characteristic that should be sought.

As can be seen from the roles exemplified in the organisation it may be that the task of maintaining certain cybersecurity elements is best delegated to whence ever the source of the original. This requires negotiation with other departmental heads as it requires allocation of their resource. They departmental heads will know if maintenance is possible within the time constraints of an individual, or if some reshuffling of work is necessary.

Part of maintenance is an element of succession planning or ensuring redundant resource to enable cover. There is a need to know who is doing what, and what happens if a key person leaves or succumbs to some illness. This becomes particularly relevant if legacy systems are involved, as one financial institution found to its cost. Part of maintenance is securing the continuity of knowledge and process. The organisation is not a collection of silos; it is a collection of interdependent interconnected systems, whether those systems are manual or technology based.

5. Intelligence

Most commercially available cybersecurity systems are reactive. Threat hunting is an intelligence based activity. There are many threats, but to carry out an informed risk assessment determining which ones warrant preventive investment, requires knowledge of the systems and its interconnections. A technical person who understands what organisational systems can be made to do and who can think like a hacker is unlikely to be on staff, but there is always hope. Technology and malefactor tactics develop rapidly, and it is no longer necessary for an attacker to have technical skill as the required software can be commissioned or acquired, ready for use, on the dark web. It is worth noting that a criminal mindset in which access is assumed, leads to a serendipitous calculation of how to monetise that access. A purposeful attack is redolent of a structured organisation, while the opposite is true for an opportunistic attack.

Intelligence is perhaps best suited to someone with a vacuum cleaner mind, who absorbs knowledge from anywhere and everywhere, analyses it internally, and makes connections. This mindset is epitomised in Six Days of the Condor (Grady, 1975), and in the 'Machine' in the television series Person of Interest (Person of Interest, 2011). It is the proactive branch of cybersecurity, and as such could also be useful in the defensive system design stage.

There are times when it useful to have social networking skills through which malefactors identified with specific skill sets can be tracked and their contacts monitored to prevent a crime as a service incident.

6. Monopoly of big tech and cybersecurity

In the move to the Cloud, organisations are outsourcing aspects of their cybersecurity, and are likely to have no access to policies and procedures or system workings of the Cloud Service Provider (CSP). This presents problems where one or more co-tenants may have the possibility to attack the underlying systems. Due to dependence and lack of alternatives, organisations may be forced to accept the assurances given by CSP.

A further concern is that public and private organisations may become the training grounds for cybersecurity resources. They will be unable to compete with the rewards offered by big tech when proficient cybersecurity experts look for their next post.

7. Conclusion

Selection to organisations is often made on the basis of outgoing personality, the ability to interact favourably in interview. It is common knowledge that all things being equal, the person that is liked more, is the one chosen. Sometimes this may not be the right choice as the best technical cyber experts may be somewhere on the autistic spectrum (Payne, et al., 2019) or operating in a different functional area of the organisation.

With the global cybersecurity workforce shortage equating to 3.5 million, it has to be asked if this is for the purely technical aspects or whether cybersecurity is assuming responsibility for all organisational failures. University courses are churning out potential pen testers, incident responders and AI security engineers, all supremely technical, but who will never be sufficient in numbers to provide all the skills required to fill the breach. The reality of cybersecurity is very different from their aspirations and expectations, and career progression not always obvious.

It is recognised that the term 'cybersecurity' may be off-putting to the non-technical. As usual, language and tone matter. Couching the consequences in relevant terms, enables the posing of questions to start discussions and enumerate solutions. Within University courses and organisational table top exercises, the questions – 'How would you do this without a computer?' or 'How essential are you to the organisation and why?', would lead to personal investment in the solutions and perhaps change the cybersecurity culture.

The inverted triangle that is the organisation need not rest solely on its tip, but could have metaphorical re-bar running throughout. Between the technical level and the Board are functions where inherent cybersecurity abilities and proclivities may be found and utilised. Education of the organisational functionaries should include elements of cybersecurity, but in audience appropriate language: the implications of consequential breach, social engineering, organisational continuity, and cyber safe understanding. In this way, cybersecurity culture will permeate the organisation, spread the load to where the continuity expertise resides, and reduce the pressing demand for the missing 3.5 million.

Cybersecurity should have a proactive service aspect, where personnel can come for advice or guidance. It is not sufficient for it to be just the reactive defence force. The answer to the skills shortage may be dependent on the definition of cybersecurity and continual acceptance of responsibility for organisational failings.

References

- Brown, G., 2013. Fourteenth Annual Sommerville Lecture: The wrong questions about cyberspace. *Military Law Review*, Volume 217, pp. 214-233.
- Cabinet Office, 2012. *Chapter 6 Business Continuity Management*. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61029/Chapter-6-Business-Continuity-Management_amends_04042012.pdf [Accessed 16 September 2021].
- Digital Guardian, 2021. *A Definition of Cyber Security*. [Online] Available at: <https://digitalguardian.com/blog/what-cyber-security> [Accessed 9 September 2021].
- Financial Reporting Council, 2016. *Guidance on the Going Concern Basis of Accounting and Reporting on Solvency and Liquidity Risks*. [Online] Available at: <https://www.frc.org.uk/getattachment/62ae3969-fe26-4def-8d25-e2acd821e7b1/Guidance-on-the-Going-Concern-REVISED-WEB-READY-2016.pdf> [Accessed 16 September 2021].
- Grady, J., 1975. *Six days of the Condor*. London: Hodder & Stoughton General Division. Group of 30, 2018. *Banking Conduct and Culture: A permanent mindset change*. [Online] Available at: https://group30.org/images/uploads/publications/aaG30_Culture2018.pdf [Accessed 16 September 2021].
- Jennings, P., 1948. Report on Resistentialism. *The spectator*, 23 April.

- Levell, J., 2017. *Rise of the Chief Neurodiversity Officer*. [Online] Available at: <https://www.neurocyber.uk/post/rise-of-the-chief-neurodiversity-officer> [Accessed 13 September 2021].
- Murphy, C. E., 2012. *Murphy's Laws Origin*. [Online] Available at: <https://militaryhumor.net/murphys-laws-origin/> [Accessed 9 September 2021].
- National Cyber Security Centre, 2021. *What is cyber security?*. [Online] Available at: <https://www.ncsc.gov.uk/section/about-ncsc/what-is-cyber-security> [Accessed 9 September 2021].
- NICCS, 2021. *Home Page*. [Online] Available at: <https://niccs.cisa.gov/> [Accessed 15 September 2021].
- Nield, D., 2020. *Researchers Just set a new record for the fastest internet speed ever*. [Online] Available at: <https://www.sciencealert.com/researchers-just-set-a-new-record-for-the-fastest-internet-speed-ever-recorded> [Accessed 26 June 2021].
- NIST, 2017. *Federal Cybersecurity Coding Structure*. [Online] Available at: <https://www.nist.gov/document/federalcybersecuritycodingstructure-v20-10-18-17docx> [Accessed 15 September 2021].
- Office of the Under Secretary of Defense for Acquisition and Sustainment, 2021. *Security Maturity Model Certification*. [Online] Available at: https://www.acq.osd.mil/cmmc/docs/CMMModelExcel_V1.02_20200318.xlsx [Accessed 15 September 2021].
- Patterson, G., 2018. *Fighting Cybercrime with Neuro Diversity*. [Online] Available at: <https://www.project-syndicate.org/commentary/neuro-diversity-cybersecurity-autism-asperger-by-gavin-patterson-2018-01?barrier=accesspaylog> [Accessed 13 September 2021].
- Payne, K.-L. et al., 2019. Is there a relationship between cyber-dependent crime, autistic-like traits and autism?. *Journal of Autism and Developmental Disorders*, Volume 49, pp. 4159-4169.
- Person of Interest*. 2011. [Film] United States of America: Kilter Films; Bad Robot Productions, Warner Brothers Television;.
- Qu, L., 2019. *99% of Fortune 500 Companies use Applicant Tracking Systems*. [Online] Available at: <https://www.jobscan.co/blog/99-percent-fortune-500-ats/> [Accessed 17 September 2021].
- Steward, M., 2021. *Compliance, Culture and Evolving Regulatory Expectations*. [Online] Available at: <https://www.fca.org.uk/news/speeches/compliance-culture-and-evolving-regulatory-expectations-mark-steward> [Accessed 16 September 2021].
- techopedia, 2021. *Finagle's Law*. [Online] Available at: <https://images.techopedia.com/definition/19342/finagles-law> [Accessed 9 September 2021].
- White, A. & Clark, B., 2017. *Blue Team Field Manual*. version 1.2 ed. Scotts Valley(California): CreateSpace Independent Publishing Platform.

Critical Systems Protection (CSP): The US Secret Service's Tactical Cyber Capability for Securing Protectees

Austin Hyman, Brian Nussbaum, Mario Bencivenga and Zachary Rizzo

College of Emergency Preparedness, Homeland Security, and Cybersecurity (CEHC),
University at Albany, Albany, NY, USA

Corresponding Author: Brian Nussbaum

padcyber@albany.edu

Abstract: This paper examines a largely unknown cybersecurity function within the United States Secret Service (USSS), that is separate and apart from its traditional cybersecurity mission in financial crime investigations. This program - Critical Systems Protection (CSP) – is instead focused on the Secret Service's protective mission and analyzes and secures the cyber environments of facilities and locations visited by the agency's protectees. Exploiting publicly available documents about CSP, this paper will document this unique organization that combines some of the highest profile issues in cybersecurity – operational and tactical level cybersecurity efforts, and the security of cyber-physical systems. Drawing on open-source information and government documents, this paper sketches out the CSP mission, its partnerships with other agencies and organizations, and some of the tools and other capabilities it has developed.

Keywords: Cybersecurity, Law Enforcement, Executive Protection, Cyber-Physical Systems

1. Introduction

In April of 2017, CBS News ran a report entitled *"Secret Service costs for Trump family protection continue to mount"* (Strickler, 2017). The article described several unusual expenses tied to protecting President Trump and his family. These expenses included travel and vehicle rentals, and other expenses that members of Congress (Strickler, 2017) and the Government Accountability Office (LaRose, 2017) were monitoring to assess their appropriateness. One particular expense stuck out enough that the CBS piece (Strickler, 2017), and one from the New Orleans Times Picayune on the same subject (LaRose, 2017), noted as unusual – the expenditure by the United States Secret Service (USSS) of \$64,000 on *"elevator services"* at Trump Tower. The idea that it is strange or inappropriate for expenditures to be made securing the physical - and especially cyber-physical systems – that high value Secret Service protectees use is to miss (or misunderstand) one of the rapidly growing and fascinating components of the Service's protective mission. The CBS piece notes *"A Secret Service official told CBS News that 'elevator inspection is one of many protective operational measures utilized by the Secret Service.'"* (Strickler, 2017) In fact, the Secret Service has rapidly expanded the cyber component of its protection mission in recent years, in large part through a program called Critical Systems Protection (CSP).

The United States Secret Service has long played an important role in federal cyber-crime and cybersecurity investigations, though with a narrower focus than an organization like the Federal Bureau of Investigation (FBI). While the FBI investigates many types of cybercrime – from espionage to sabotage to fraud to data theft – the United States Secret Service has focused predominantly on cybercrimes targeting the financial sector and industry. This owes to the agency's foundational focus on economic crimes (Bumgarner et al 2017), which - while lesser known than their high-profile protective mission - shaped their approach to cybercrime.

The Secret Service has run Electronic Crimes Task Forces (ECTFs) since the mid 1990s (NLECTC, 2000) in conjunction with other Federal partners, as well as state and local agencies (USSS, 2021); however, these task forces have traditionally hewn close to the agency's focus on financial cybercrime. In 2020, the ECTFs were merged with the Secret Service's Financial Crime Task Forces (FCTFs), to create a new series of Cyber Fraud Task Forces (CFTFs) (USSS, 2020). As of mid 2020, there were 42 CFTFs in the United States, and international ones in London and Rome (USSS, 2020). The Service reports that *"Together with our partners, the CFTFs stand ready to combat the full range of cyber-enabled financial crimes."* (USSS, 2020)

In addition to the focus on financial crime, there have been several other themes that have defined Secret Services cyber efforts. Partnerships have been key to their building cyber investigation capabilities. Sometimes these have been partnerships with state and local government agencies through various task forces, but there have also been partnerships with academic institutions like Carnegie Mellon (on cybersecurity) and the University of Tulsa (on mobile phone security) (Lemieux, 2011). The Secret Service has also tried to build up the capabilities of state and local partners, through efforts like the National Computer Forensics Institute (NCFI) –

“the nation’s only federally funded training center dedicated to instructing state and local law enforcement officers, prosecutors, and judges in cybercrime investigations.” (Gaskew, 2019)

2. The CSP Mission

2.1 Program and Function

The Critical Systems Protection program is a component of the United States Secret Service’s protective mission that utilizes specially trained agents to identify, assess and mitigate cyber risks (GAO, 2020; DHS, 2020; GAO, 2014; Clancy, 2015). The program is responsible for evaluating cybersecurity threats to people, places and events protected by the USSS (DHS 2017; Clancy, 2015; Cybersecurity, 2013). The goal of the CSP is to strengthen the USSS’s protective operations by analyzing how information networks may impact the physical security plan or cause physical harm to the protectee (CERT, 2008; DHS, 2008). Additionally, the CSP is responsible for developing practical mitigation strategies that minimize the risk posed by these cyber impacts and prevent adverse impacts on protection (CERT, 2008, DHS, 2008). That is, the program leverages many of the skills and techniques that Secret Service agents and analysts bring to bear on traditional cybercrime, but does so not for financial criminal investigations, but rather to help protect the government officials and other high-profile targets that the Service is charged with protecting.

Operationally, the CSP relies on the expertise of its agents working in the Office of Investigations and external partnerships to secure the cyberspace surrounding protectees. The capability of investigative agents, especially those working on ECTFs or now CTFs, to conduct network monitoring and tracking are used in protective operations as a part of CSP advances (Goldstein, 2016; GAO, 2014). The CSP also works with partners inside and outside of the Federal government to develop tools, conduct research and increase capabilities (Sandia, 2019; DHS, 2013; CERT, 2008; SEI, 2002). In practice, the CSP is coordinated by agents in the Office of Investigations who are directed by a lead advance agent from the Protective Division. (GAO, 2020; DHS, 2011). The CSP typically deploys in advance of a protectee visit and coordinates with investigative agents from local field offices to conduct assessments and develop security plans (Senate Appropriations, 2017; Maui, 2017; Cybersecurity, 2013). During an event, dedicated CSP agents, and investigative staff are pulled from their normal duties, to monitor and respond to potential cyber threats to the operation or event, and to the protectee (GAO, 2020; DHS, 2019; Weiqun, 2017; Lowery, 2014).

Budget of Operational Mission Support (OMS) Initiatives (US Secret Service)						
	FY 16 Revised Enacted	FY 2017 Annualized CR	FY 2018 Request	FY 2019 Enacted	FY 2020 Enacted	FY 2021 Presidents Budget
Cyber Protection Activities Budget	\$6.3 million	\$6.3 million	\$6.3 million	\$6.8 million	\$7.002 million	\$8.2 million
Source	(DHS, 2018)	(DHS, 2018)	(DHS, 2018)	(Treasury, 2021)	(Treasury, 2021)	(Treasury, 2021)

2.2 Cyber Assessment of Protectee Visit Sites

The primary duty of the CSP is to conduct cyber assessments at venues and events where a protectee will spend time. Members of the CSP deploy in advance of protectee visits to conduct systematic audits and technical assessments of the sites where the protectee will be, as well as of the critical infrastructure and utilities that support the visit (Stegon, 2016; GAO, 2016; Clancy, 2015; DHS, 2011; Harlow, 2011). Advance assessments may begin as early as nine months pre-event to ensure that the CSP agents have identified and mitigated all potential cyber threats to the protective mission (Weiqun, 2017). CSP advance teams perform strategic assessments of the host city at the macro level to examine critical infrastructure and IT networks that may impact physical security. (DHS, 2019; Weiqun, 2017; Stegon, 2016; Clancy, 2015; DHS, 2011). According to Congressional testimony by former USSS Director Mark Sullivan, these assessments “*identify which computer networks, process-control systems, or remotely controlled devices could, if compromised, impact an operational security plan*” (Stegon, 2016). Agents also integrate and apply the latest threat intelligence to their operational security plans, collaborating with other members of the U.S. Intelligence Community to understand the current cyber threat landscape (GAO, 2014; GAO, 2020; Weiqun, 2017; USSS, 2013).

At the micro level, the CSP will assess the site where a protectee is visiting to mitigate risks to the actual facility. Advance teams evaluate buildings’ network infrastructure and IT systems to identify potential vulnerabilities

(Goldstein, 2016; GAO, 2020; Harlow, 2011). The methodologies employed by the CSP prioritize systems that may directly affect the safety of the protectee if disrupted or destroyed (GAO, 2020; Maui, 2017; GAO, 2014; CERT, 2008). Once vulnerabilities are identified, the CSP works with facility owners and operators to mitigate risks and protect their information systems (Goldstein, 2016; Maui, 2017). During the protectee's visit, CSP monitors networks and devices that could be compromised in-real time for potential breaches of security (GAO, 2020; DHS, 2019; DHS, 2013). At protected venues, agents ensure failsafe security controls are in place and are prepared to act in defense of IT networks in the event of malicious cyber activity (Weiqun, 2017; DHS, 2019).

Number of Advances Per Fiscal Year					
FY	FY 13	FY 14	FY 15	FY 16	FY 17
Number of Advances Per Fiscal Year	186	230	241	263	264
Source	(USSS Annual Report, 2013)	(USSS Annual Report, 2014)	(USSS Annual Report, 2015)	(USSS Annual Report, 2016)	(USSS Annual Report, 2017)

2.3 Cyber Assessments for NSSEs

In addition to protecting people, the CSP is tasked with protecting events designated as National Special Security Events (NSSEs). NSSEs are events designated by the Secretary of DHS or President of the United States that are of national significance and may be a target for criminal or terrorist activity (CRS, 2021). The USSS is the lead federal agency for developing and implementing security operations at NSSEs, of which cybersecurity is a critical aspect (DHS, 2019). The CSP has been involved with securing NSSEs since its inception, working to secure the 2002 Winter Olympics in Salt Lake City, Utah and 2002 Superbowl XXXVI in New Orleans, Louisiana (Lowery, 2014; SEI, 2003). The CSP scales its operations up to meet the needs of NSSEs, which are some of the most complex and logistically complicated protective operations undertaken by the USSS (DHS, 2019; Lowery, 2014). CSP employs very similar methodologies at NSSEs as it does for protective visits, working to identify and mitigate cyber risk at both the macro and micro levels (DHS, 2019; Stegon, 2016; Cybersecurity, 2013). The CSP has deployed to several NSSEs, including Presidential inaugurations, major international summits held in the U.S., major sporting events, and Presidential nominating conventions (CRS, 2021; Senate Appropriations, 2016; Weiqun, 2017; CERT, 2008).

2.4 Support to Other Missions

The CSP utilizes its expertise in cyber protection and risk mitigation to secure critical infrastructure in support of greater homeland security objectives (A. DHS, 2020). As a part of this effort, agents from CSP collaborate with critical infrastructure owners and operators to secure their systems from cyber threats (US Senate, 2013). CSP agents have conducted trainings with critical infrastructure stakeholders and local governments on the awareness of real-time network activity and the identification of malicious activity (DHS, 2013; DHS, 2008). In addition to this, CSP personnel have been requested to assist in assessing cyber threats to specific critical infrastructure facilities (US Senate, 2013). These assessments identify the critical components of a facility's cyber infrastructure and detail any potential vulnerabilities to them, as well as provide recommendations for mitigating those threats (US Senate, 2013). While details on specific CSP assessments of critical infrastructure operators are limited, its agents did conduct an assessment with a maritime port stakeholder in the Houston, Texas area and have engaged with similar stakeholders in California (US Senate, 2013).

3. Capabilities and Tools

3.1 Capabilities

While the full extent of CSP's capabilities is hard to judge based on publicly available information, the program has launched many hundreds of advance missions and has secured numerous types of events. In addition to the broader role of providing a cyber component to the protective mission, CSP has a number of narrower functions of importance. The CSP engages in training, software development, as well as representing the Secret Service at various types of events. Examples of such activities include:

3.1.1 Training

- Conducted training on the awareness of real-time network activity through traffic analysis and the identification of malicious network activity (DHS, 2013).

- Expanded the CSP program to 20 agents and increased capabilities to include a network operations center, R+D training and operation units (DHS, 2013).

3.1.2 Representation

- Represented the Secret Service and served as the sole presenter from the United States, at INTERPOL's request, in assisting the Government of Qatar in preparing for the 2020 FIFA World Cup (DHS, 2015).
- Collaborated with the Intelligence Community and private sector to support the State of the Union Address, the George W. Bush Presidential Center dedication, and the 68th United Nations General Assembly (DHS, 2013).

3.1.3 Technological Developments

- Developed and deployed a remote sensing platform with the Software Engineering Institute at Carnegie Mellon University (DHS, 2013).
- Conducted training on the awareness of real-time network activity through traffic analysis and the identification of malicious network activity (DHS, 2013)
- Successfully deployed the first full-time, USSS/CSP Liaison in partnership with the Executive Office of the President and the White House Communications Agency. This collaboration directly joins key forces in identifying cyber threats targeting USSS protectees and venues. This effort has enriched the threat information sharing platform and enhanced the cyber-security posture (DHS, 2016).
- Enhanced the CSP Program's Computer Network Defense mission by implementing sensors on a critical network managing energy delivery; thus, allowing the Service to proactively address cyber threats to its protective mission (DHS, 2016).

Research and Development Budget						
	FY 16 Enacted	FY 2017 Enacted	FY 2018 Enacted	FY 2019 Enacted	FY 2020 Enacted	FY 2021 Presidents Budget
CERT Program	\$250,000	\$250,000	\$250,000	\$250,000	\$1,750,000	\$250,000
Source	(DHS, 2018)	(DHS, 2018)	(DHS, 2018)	(Treasury, 2021)	(Treasury, 2021)	(Treasury, 2021)

3.2 Tools

Analysis of publicly available documents provides insight into the capabilities of CSP and the tools they use to conduct their mission. The CSP is currently using at least two major defense tools created by the Computer Emergency Response Team (CERT) at Carnegie Mellon University. These tools include the Kaleidoscope network defense platform and the Compass protective advance expert system application - formerly known as FlipBook (Treasury, 2021). *"CERT also developed the Bank Note Processing System (BNPS), which is an application that increased the efficiency of processing counterfeit bank notes in Secret Service field offices."* (A. DHS, 2020) CERT is not the only organization designing special tools and applications for CSP. The Tamizar tool, initially designed by Sandia National Lab for the U.S. Army, has become the primary network forensics tool for CSP (Sandia, 2019). The USSS relies on Sandia capabilities and expertise as part of a long standing and successful partnership. (Sandia, 2019). CSP has partner with other organizations to develop their tools, according to the US Secret Service yearly report in 2015 *"Partnered with two federally funded research and development centers, in developing customized tools to enrich the CSP Program's Network Traffic Analysis."* (DHS, 2015).

4. Partnerships

4.1 University, Research, and Analytical Partnerships

The Secret Service has recognized that partnerships are an integral part of both its investigative and protective missions (Maui, 2017). At the turn of the century, when the rise of cybercrime became more prolific, the USSS reaffirmed its commitment to work together and leverage relationships through its Critical Systems Protection Program (CSP) (Maui, 2017). The first of these partnerships began in 2001, with the collaboration of the U.S Secret Service and CERT at Carnegie Mellon (USSS, 2008). Through this partnership, the Insider Threat Study (ITS) was produced (USSS, 2008). Employing many of the same research methods as the Exceptional Case Study Project conducted by the USSS in 1977, the ITS focused on identifying information that may have been noticeable before an incident *"from both a behavioral and technical perspective"* (Harlow, 2001).

However, the collaboration between CERT at Carnegie Mellon and the U.S Secret Service would progress further, resulting in the creation of the Critical Systems Protection program. This initiative was designed to analyze the relationship between critical information networks and physical protection activities (SEI, 2003). Often used to support law enforcement, CSP was developed through the support of events such as the 2004 Olympic games, 2005 Presidential inauguration, and 2005 Olympic Games (CERT, 2008). When CSP is employed, often building owners and operators are worked with to ensure a building's cybersecurity is satisfactory (Goldstein, 2016).

Within two years since the start of this collaboration, Secret Service electronic crimes experts testified that this effort was “truly groundbreaking” in its ability to combine both the physical and cyber aspects of the Secret Service’s protective mission (USSS, 2003). In 2011, CSP once again utilized the resources of CERT, as well as adding both Carnegie Mellon University’s CyLab and Center for Sensed Critical Infrastructure Research (CenSCIR) to their collaborative effort (DHS, 2011). This partnership sought to create additional resources to further monitor critical infrastructure, systems, as well as a wide range of physical, cyber, and environmental threats (DHS, 2011).

However, Carnegie Mellon University has not been the only long-term partnership CSP has formed. Sandia National Laboratories has been a key contributor to the program for both its software development and cyber protection capabilities. In addition to the Tamizar software, Sandia Laboratory provided additional cyber support to the Secret Service (Sandia, 2019). In 2019, Sandia Laboratory was called upon by the USSS to provide “*onsite and remote cyber analysts*” for President Donald Trump’s State of the Union address, being one of only two agencies tasked with this mission. (Sandia, 2019). Sandia Laboratory has called this relationship “a long-standing and successful partnership.” (Sandia, 2019).

4.2 State, Local, Tribal, and Territorial (SLTT) Partnerships

Yet not all CSP’s partnerships are created with the intent of developing new tools or or analytic methodologies. Some, like the collaboration between the Maui County Police Department and USSS, are formed to leverage each other’s assets and capabilities. This example of a federal-local partnership was formed in 2017, with the USSS and Maui County Police department signing an intergovernmental agreement to enable joint operations and help organize the funding of such collaborations (Maui, 2017). The need for this agreement was born from the increase in “*network intrusion, malicious software, and unauthorized use of personal identification information*” in the region (Maui, 2017). To mitigate and deter these activities, the Secret Service provided a small team of specialized agents trained in network intrusions as well as Critical Systems Protection (Maui, 2017).

Beyond the level of counties, municipalities have also partnered with the CSP. The City of Beverly Hills is one of these partners (Beverly Hills, 2010). In 2010, the City of Beverly Hills Hazard Mitigation Action plan sought to conduct an assessment and analysis of critical infrastructure areas (Beverly Hills, 2010). The goal of this assessment was to observe how areas interfaced with physical and cyber components would be affected in a potential terrorist attack (Beverly Hills, 2010). To conduct this assessment, the city recommended utilizing the Secret Services’ CSPI (Beverly Hills, 2010). In 2016, a similar assessment was recommended (Beverly Hills, 2016). This time, the project focused on city facilities, including a metro line assessment. Once again, the use of CSPI was employed (Beverly Hills, 2016).

5. Conclusion

The Critical System Protection program within the United States Secret Service is a pioneering, and largely unknown program in the world of cybersecurity. This despite it being at the nexus of two of the most exciting and emerging cybersecurity issues – the first is that of securing cyber-physical systems, and the second is the idea that there can be some kind of “tactical” or “operational” cybersecurity that occurs in real-time as opposed to focusing just on cyber hygiene and the implementation of frameworks of controls. Without much fanfare, the Secret Service conceived of, built, funded, and operationalized just such a tactical cybersecurity function with a focus on operational technology and cyber-physical systems to bolster the security of its protectees. In most of cybersecurity, the asset being protected is data. In the world of Critical Systems Protection, their crown jewels really are High Value Targets.

References

- Bumgarner, J. Crawford, C. Burns, R. (2017) Federal Law Enforcement: A Primer. Carolina Academic Press.
- CERT. (2008). *CERT Research Annual Report 2008*. Carnegie Mellon, Software Engineering Institute. Retrieved from: https://resources.sei.cmu.edu/asset_files/CERTResearchReport/2008_013_001_54241.pdf

- City of Beverly Hills. (2010). Hazard Mitigation Action Plan: 2010-2015. Retrieved from http://www.beverlyhills.org/cbhfiles/storage/files/filebank/10286--9a_HazardMitigationPlan%201152011.pdf
- City of Beverly Hills. (2016). Hazard Mitigation Action Plan: 2016-2021. Retrieved from [http://www.beverlyhills.org/cbhfiles/storage/files/1388644851974314648/BeverlyHillsFinalHMP\(002\).pdf](http://www.beverlyhills.org/cbhfiles/storage/files/1388644851974314648/BeverlyHillsFinalHMP(002).pdf)
- Congressional Research Service. (2021). *National Special Security Events: Fact Sheet*. Prepared for Members and Committees of Congress, CRS Report R43522. Retrieved from <https://fas.org/sgp/crs/homesecc/R43522.pdf>
- Department of the Treasury, United States Secret Service. (2021). *Congressional Budget Justification and Annual Performance Plan and Report*. FY 2021. Retrieved from <https://home.treasury.gov/system/files/266/15.-USSS-FY-2021-CJ.pdf>
- Gaskew, B. (2019). Reader's Guide to Understanding the US Cyber Enforcement Architecture and Budget. Third Way. Retrieved from <https://thirdway.imgix.net/pdfs/override/Memo-Readers-Guide-Cyber-Budget.pdf>
- Goldstein, P. (2016). GITEC 2016: Cybersecurity complexity will Increase; Secret Service official says. *FedTech Magazine*. Retrieved <https://fedtechmagazine.com/article/2016/04/gitec-2016-cybersecurity-complexity-will-increase-secret-service-official-says>
- Harlow, R. (2011). *Two Missions, One Secret Service: The Value of The Investigative Mission* (pp. 1-85). [Master's thesis, Naval Postgraduate School]. Monterey, California: Naval Postgraduate School. Retrieved from <https://apps.dtic.mil/dtic/tr/fulltext/u2/a552250.pdf>
- Innovation Law Lab, et al., v. Kirstjen Nielsen, et al., 3:19-cv-00807-RS. (United States District Court Northern District of California). 2019. Retrieved from <https://egov.gov/office-locator/#/asyresults/CA/California>
- LaRose, G. (2017) Trump family Secret Service costs piling up: report. *Times-Picayune*. April 14. Retrieved from https://www.nola.com/news/politics/article_3c5ab78a-c293-5df4-b88f-afab14f11961.html
- Lemieux, F. (2011). Investigating cyber security threats: Exploring national security and law enforcement perspectives. *2011 Developing Cyber Security Synergy*, 63.
- Lowery, E. W. (2014). *Closing the cyber gap: integrating cross-government cyber capabilities to support the DHS cyber security mission*. [Master's thesis, Naval Postgraduate School]. Monterey, California: Naval Postgraduate School. Retrieved from <https://www.hsdl.org/?view&did=762425>
- Maui County Government. (2017) A Bill For An Ordinance Authorizing The Mayor of The County of Maui to Enter Into An Intergovernmental Agreement With The United States Secret Service (USSS) Regarding Payment of Expenses from Joint Operations Between USSS and The Maui Police Department, Bill #81, 2017. Retrieved from <https://www.maui-county.gov/DocumentCenter/View/110670/Ord-4772>
- National Law Enforcement and Corrections Technology Center (NLECTC). (2000). *Catching the Cyber Crook*. Retrieved from <https://www.ojp.gov/pdffiles1/nij/nlectc/211823.pdf>
- Report on 2017 Department of Homeland Security Appropriations Bill, 114th Congress, U.S. Senate, Committee on Appropriations (2016). Retrieved from <https://www.congress.gov/114/crpt/srpt264/CRPT-114srpt264.pdf>
- Sandia National Laboratories. (2019). *Labs Accomplishments: National Security Programs*. Sandia LABNEWS. Retrieved from https://www.sandia.gov/news/publications/lab_accomplishments/articles/2019/national_security_programs.html
- Software Engineering Institute. (2003). *SEI Annual Report FY 2002*. Carnegie Mellon, Software Engineering Institute. Retrieved from <http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=30178>
- Statement of Director Joseph P. Clancy United States Secret Service Department of Homeland Security, 114th Congress, U.S. House of Representatives, Before the Committee on Oversight and Government Reform. (2015). Joseph P. Clancy. Retrieved <https://docs.house.gov/meetings/GO/GO00/20150324/103244/HHRG-114-GO00-Wstate-ClancyJ-20150324.pdf>
- Stegon, D. (2016). Secret Service Director Testifies on FY13 Budget. *FedScoop*. Retrieved from <https://www.fedscoop.com/secret-service-director-testifies-on-fy13-budget-1/>
- Strickler, L. (2017) Secret Service costs for Trump family protection continue to mount. *CBS News*. April 14. Retrieved from <https://www.cbsnews.com/news/secret-service-costs-for-trump-family-protection-continue-to-mount/>
- United States Department of Homeland Security, United States Secret Service. (2008). *United States Secret Service Strategic Plan* (FY 2008-2013). Retrieved from <https://www.hsdl.org/?view&did=487083>
- United States Department of Homeland Security, United States Secret Service. (2011). *United States Secret Service Fiscal Year 2011 Annual Report*. Retrieved from <https://www.secretservice.gov/>
- United States Department of Homeland Security, United States Secret Service. (2013). *United States Secret Service Fiscal Year 2013 Annual Report*. Retrieved from https://www.secretservice.gov/sites/default/files/reports/2021-03/USSS_FY2013AR.pdf
- United States Department of Homeland Security, United States Secret Service. (2014). *United States Secret Service Fiscal Year 2014 Annual Report*. Retrieved from https://www.secretservice.gov/sites/default/files/reports/2021-03/USSS_FY2014AR.pdf
- United States Department of Homeland Security, United States Secret Service (2015). *Fiscal Year 2015 Annual Report*. Retrieved from <https://www.secretservice.gov/sites/default/files/reports/2021-03/USSSAnnualReportFY2015.pdf>
- United States Department of Homeland Security, United States Secret Service (2016). *Fiscal Year 2016 Annual Report*. Retrieved from https://www.secretservice.gov/sites/default/files/reports/2021-03/USSS_FY2016AR.pdf
- United States Department of Homeland Security, United States Secret Service (2017). *Fiscal Year 2017 Annual Report*. Retrieved from https://www.secretservice.gov/sites/default/files/reports/2021-03/CMR-2017_Annual_Report_online.pdf

- United States Department of Homeland Security, United States Secret Service. (2018). *U.S. Secret Service Budget Overview*. Retrieved from <https://www.dhs.gov/sites/default/files/publications/USSS%20FY18%20Budget.pdf>
- United States Department of Homeland Security, United States Secret Service. (2020b). *Office of Investigations Strategy Fiscal Year 2021-2027*. Retrieved from <https://www.secretservice.gov/sites/default/files/reports/2021-01/inv-strategy-fy21-27.pdf>
- United States Department of Homeland Security. (2019). *U.S. Secret Service Budget Overview: Fiscal Year 2019 Congressional Justification*. Retrieved from https://www.dhs.gov/sites/default/files/publications/19_0318_MGMT_CBJ-Secret-Service_0.pdf
- United States Department of Homeland Security. (2020a). *U.S. Secret Service Budget Overview: Fiscal Year 2020 Congressional Justification*. Retrieved from https://www.dhs.gov/sites/default/files/publications/19_0318_MGMT_CBJ-Secret-Service_0.pdf
- United States Government Accountability Office. (2014). *Critical Infrastructure Protection: DHS Needs to Enhance Efforts to Address Port Cybersecurity* (GAO-14-459). Report to the Chairman, Committee on Commerce, Science, and Transportation, U.S. Senate. Retrieved from <https://www.gao.gov/assets/gao-14-459.pdf>
- United States Government Accountability Office. (2016). *U.S. Secret Service Data Analyses Could Better Inform the Domestic Field Office Structure* (GAO-16-288). Report to Congressional Requesters. Retrieved from <https://www.gao.gov/assets/gao-16-288.pdf>
- United States Government Accountability Office. (2020). *Secret Service Investigation Operations Confer Benefits, but Additional Actions Are Needed to Prioritize Resources* (GAO-20-239). Report to the Committee on Homeland Security and Governmental Affairs, U.S. Senate. Retrieved from <https://www.gao.gov/assets/gao-20-239.pdf>
- United States Senate, Committee on Appropriations (2013) *Cybersecurity: Preparing For and Responding To The Enduring Threat*. 113th Congress, Committee on Appropriations. Retrieved from <https://www.govinfo.gov/content/pkg/CHRG-113shrg81526/html/CHRG-113shrg81526.htm>
- USSS. United States Secret Service (2003). *U.S. Secret Service Electronic Crimes Experts Address House Subcommittee on "Fighting Fraud: Improving Information Security"*. U.S. Secret Service Media Relations. Retrieved <https://www.secretservice.gov/press/releases/2003/04/us-secret-service-electronic-crimes-experts-address-house-subcommittee>
- USSS. United States Secret Service (USSS) Director's Update newsletter (2008) produced from January 1, 2001 to August 10, 2004. Retrieved from https://www.governmentattic.org/docs/USSSDirectorsUpdate_2001-2007.pdf
- USSS. United States Secret Service. (2020) *Secret Service Announces the Creation of the Cyber Fraud Task Force*. Retrieved from <https://www.secretservice.gov/newsroom/releases/2020/07/secret-service-announces-creation-cyber-fraud-task-force>
- USSS. United States Secret Service. (2021) *ECTF and FCTF*. Retrieved from <https://www.secretservice.gov/contact/ectf-fctf>
- Weiqun, Z. (2017). What does it take to (cyber)secure the presidential inauguration? *Federal Times*. Retrieved from <https://www.federaltimes.com/federal-times/2017/01/09/what-does-it-take-to-cyber-secure-the-presidential-inauguration/>

WFH, not WTH? The security challenges of working-from-home

Neal Kushwaha¹, Piret Pernik² and Bruce W. Watson³

¹IMPENDO Inc., Ottawa, Canada

²Strategy Branch, NATO CCDCOE, Tallinn, Estonia

³IP Blox and IMPENDO Inc., Eindhoven Netherlands and Ottawa, Canada

neal@impendo.com

piret.pernik@ccdcoe.org

bruce@ip-blox.com and bruce@impendo.com

Abstract: Under the coronavirus pandemic, governments and corporations around the world have adopted a work-from-home (WFH) mode of operations to continue governing and operating. Over a two year into the COVID-19 pandemic, many of us continue to work from home and a large majority have few plans to return to the office. Early on, governments and companies scrambled to increase Virtual Private Network (VPN) licenses and bandwidth capacities to take on the additional user load at a technical level. This allowed a near seamless continua of communications for common government unclassified information and corporate sensitive information of non-national interest using commercial software encryption. But what about information of national interest? A smaller number of individuals in key government departments, sometimes under staff rotations, continued to work in the office to serve these needs. Within weeks, government departments began deploying assets to access classified Secret systems from home. This paper discusses the WFH use of classified (e.g., Secret) IT systems while considering multiple security areas (physical, operational, personnel, IT, communication, and electromagnetic and radio-frequency emission) with focus on insider threats and foreign state actors, to describe the impact to the WFH public servant, the citizens, and the government. It describes the severe security challenges and risks governments have accepted under the pandemic, raising the question “*what the heck* were governments thinking?”

Keywords: work from home, government, insider threats, foreign state actors, national interest

1. Introduction

The COVID-19s pandemic has reshaped the way businesses and governments operate. Even with the ability to do it remotely, many companies have been permanently shut. The *work from home* (WFH) method of conducting business operations *wherever possible* has become the new way of working for many, including governments. Never has there been an event to drive a global WFH during the information age. While some research has been performed under the notion of working from home (Giri 2022) (Carvin 2021) (Vériter, et al. 2021), the research into the negative impact to national interests due to classified WFH has not been described.

Each government designs, manages and implements policies to categorise information. These categories closely align to their foreign counterparts, but still remain different. In Canada, it is categorised as follows (Treasury Board of Canada Secretariat 2019a):

1. Unclassified: normally information that has been declassified, is of low value, and has negligible-to-low impact if released; the information may already be publicly available; in some states, the term also means the information referenced is “not yet classified” while in others, it includes the information classified for internal use.
2. Non-national interest: in various states, this information is still considered unclassified but includes other designations to represent its sensitivity; across various States, the designation is represented by terms such as RESTRICTED (e.g. USA), PROTECTED (e.g. Canada), OFFICIAL, INTERNE, SBU (*sensitive but unclassified*), and others; non-national interest information is of low impact to the state but can be of varying degrees of impact to an individual including physical injury, kidnapping, and loss of life.
3. National interest: this category is commonly subcategorised into three in Canada (Treasury Board of Canada Secretariat 2019a) or four sub-containers at NATO and other states (Estonian National Security Authority n.d.); while each sub-container’s classification label is not strictly agreed upon by all governments, for purpose of this paper, the level of general understanding described below is adequate.
 - RESTRICTED;
 - CONFIDENTIAL: release of information is likely to cause injury to the national interest;
 - SECRET: release of information is likely to cause grave injury to the national interest; and
 - TOP SECRET: release of information is likely to cause exceptionally grave injury to the national interest.

Government civil servants compose and share various documents that are restricted on access by legislation even while in draft. Governments maintain procedures, regulation, and laws within which national interest information is created, transmitted, stored, and destroyed at a common standard across all its branches. Other sub-controls exist to further limit the exposure of the information impacting national interest, for example dissemination controls (e.g., TOP SECRET//CANADIAN EYES ONLY, or TS//CEO in short form).

Under Canadian legislation, lead security agencies (Treasury Board of Canada Secretariat 2019b) of the federal government provide guidance that national interest information should not reside or be transmitted in *plain text* (or *in the clear*) on publicly accessible networks and that each of the systems that carry national interest information must adhere to increasingly strict security controls as the injury to the national interest increases (Canadian Centre for Cyber Security 2018).

The content creator is commonly responsible to classify the information and that is one of the many reasons why information is incorrectly classified. While domestic laws exist to manage the release of national interest information, in some states including Canada, governments must prove intent in order to convict. Since classified information is only available in *plain text* on government physically isolated systems,¹ the greatest threat to the confidentiality, integrity, and access remains with the users, who are considered insider threats. Insider threats can release information (1) without intent, (2) intentionally without malice, and (3) with malicious intent.

Security professionals with traditional experience come together with technological experienced security professionals to help mitigate and safeguard the threat and risks that states face. To facilitate reading the remainder of the paper, we define these security domains² as follows.

1. Non-technological security domains:
 - Physical security (PhySec): the practise of securing people and assets from unauthorised access and natural threats that result in a breach of confidentiality, damage, or loss.
 - Operational security (OPSEC): the assessment and protection of publicly available information and observable behaviours, that when collectively analysed by an adversary, disclose compromising information or information that was intended to remain undisclosed.
 - Personnel security (PERSEC): the evaluation of an individual's reliability, loyalty, and trustworthiness; it also discloses how the individual can be compromised, discredited, and even eliminated.
2. Technological security domains:
 - IT security (ITSEC): a discipline concerned with protection of information technology using tools, policies, procedures and more from unauthorised access.
 - Communications security (COMSEC): a discipline that ensures secure electronic transmittal of information, while preventing unauthorised access using cryptography.
 - Emission security (EMSEC): the practise to identify and reduce the interception of signal emissions from technological equipment.

In WFH conditions, the challenges in each of these categories are as follows:

1. PhySec: Physical security policies (how to define security zones and controls at home).
2. OPSEC: Operational security policies (difficult for the security office to oversee users working from home).
3. PERSEC: Personnel security challenges (new people at home potentially having access or awareness of information and possibly how to access the information).
4. ITSEC: IT security considerations (exponential growth in sources of entry, and thus attack surface).
5. COMSEC: Communication security doctrine (complications in granting access to COMSEC equipment from home).
6. EMSEC: Electromagnetic and radio-frequency emission security concerns (greater threat from uncontrolled homes versus well-spaced and protected government campuses).
7. Insider threats: Users can likely bring other devices into the frame such as cameras to capture screens and share information through other means that are out of the control of the government.

¹ Isolated and private networks can still be globally accessible.

² These security domains are very intermingled, e.g., communications security professionals consider threats and risks related to physical security, operations security, personnel security, and emission security.

8. Foreign state actors: With many foreign state actors unable to travel to countries and within countries, how might they use other means to exploit individuals (sabotage) and exfiltrate information (espionage) and continue with their mission?

Under the pandemic, the norms of safeguarding restricted information have been relaxed, increasing the threat of domestic and foreign espionage and sabotage. Phone calls that were *never* made on open public telecom infrastructures are now being held on public mobile phones in order to continue governing, for example under the US NETCOM, the US Army will deploy access to SIPRNet (classified SECRET) for over 2000 users (Eversden 2020). In some cases, tactical conversations with minimal cover-time may be justified due to low risk, however, strategic ones with greater cover-time and greater risk remain much more difficult to justify. The same issues apply to documents of national interest over public transports (e.g., Internet).

Our paper describes the various risks that are being accepted by governments, as they authorise national interest information to be created, shared, stored, and destroyed outside of their common controls at home.

The following sections will outline these non-technological and technological risks by illustrating them with commonly used WFH scenarios.

2. Non-technological security

When we think of government facilities, swipe cards, security guards, bollards, cameras, and more come to mind. Government offices come in a varying degree of styles and locations. Regardless of where the building is situated, it is an obvious source of Physical Security (PhySec), Operations Security (OPSEC), and Personnel Security (PERSEC) information for observers to ascertain.

Simple matters can be understood by simply watching the resources entering and exiting the campus or building(s). Their mode of transportation to and from may include for example, public transit or vehicle make and model, which can help ascertain one's income and if followed home, can help identify where they live, the type of home in which they live, their dependents, and other habits they may have.

Facility data that can be gathered from simply watching the resources enter a campus or building(s) include working hours of the resource, general working hours of the facility, ingress and egress points of the facility, and specific entry points for office workers, maintenance resources, and even deliveries. Based on the type of site or department hosted in the facility, the facility clearance level may also be ascertained and thereby the likely level of clearance granted to those who enter and exit the facility on a regular basis. Resources entering the campus or facility often must present a badge and most resources do not pocket or hide their badges while off campus. Some badges indicate clearance levels or status of the resource (through colour coding or possibly simply letters), thereby offering further data to the adversary watching from a far.

Resources arriving in seemingly professional, or business attire may indicate a more senior level resource, regular casually dressed resources may indicate a general civil servant, whereas a resource in uniform may indicate defence related or possibly facilities management resources.

Once an adversary identifies a resource as a valuable asset, they can gather much more specific data through targeting tradecraft.

In the few bullets above, the surveillance at the external PhySec layer of the facility ties closely to OPSEC and PERSEC. Open and unoccupied land surrounding a building help the departmental security team by offering time to intercept a perpetrator that may be approaching, trespassing, or climbing over a distant perimeter fence, while layers of fencing can slow the perpetrator. The private roads on the campus help control access and allow for the security team to evict unauthorised or suspicious vehicles, and potential observers (in vehicles or as pedestrians).

Considering only the three security principles above, if a civil servant with security clearance (granted to access and compose classified materials) were to *work from home* under the pandemic after only having worked at the office for years, they would likely not have access to equipment or methods to connect to their government departmental services to work on national interest matters from home. The civil servant may not have a portable

device of any kind to work from home because their normal workplace has always been in a secure government facility authorised for such classified workloads.

2.1 Scenario 1: WFH with authorised government equipment

Under normal working circumstances, in order to consume and/or create classified materials, the civil servant would need to travel to work to use classified compute devices. In this scenario, under the pandemic restrictions, the civil servant would need to (1) travel to the office to pickup authorised telework devices (or classified WFH kit) to be used from their residence, or (2) have authorised personnel from the office travel to the civil servant's residence to deliver and/or install telework devices (or classified WFH kit). In the latter two cases, the government department must already have the technological capability in place to offer the civil servant access to the national interest collaboration environment through secure telework communication capacities.

In any case, if the civil servant was already a target by a domestically located foreign threat actor, their behaviour at home is likely already known and if they are not already a target, working from home during the pandemic might not be a grave physical security risk. Or is it?

If the civil servant has been granted equipment to securely and remotely access and collaborate on matters of national interest (e.g., equivalent to classified SECRET such as the US and their WFH deployed access to SIPRNet), many questions arise.

A skilled PhySec resource should be consulted to ascertain the risks associated with the residence. The type of residence, e.g., an apartment building, a semi-detached, or single home, introduces different types of considerations and concerns. Each will most likely require further investigation into the neighboring residents and their affiliations, but more so for the apartment buildings and semi-detached buildings.

The PhySec resource would also assess the physical construction of the residence and measure them to a government facility standard required to host classified assets. For example: wall and floor construction using reinforced concrete versus wood and/or poly structure, steel doors and steel door frames attached to steel studded walls, and window frames along with windows lined with film to limit and deter break entries.

Government facilities are constantly monitored by an authorised group of individuals skilled at handling incidents and intercepting individuals who may be lurking or attempting to breach the perimeter, long before they reach the facility. The PhySec resource assessing the residence would likely assess the security controls surrounding the home, such as sensors and their coverage of all access points, wired or wireless sensors and battery maintenance, and intrusion alarm systems. They would additionally assess whether such systems are normally armed even when the residence is occupied, who has access to enable or disable the alarm system, the presence of camera systems and where they are pointed, as well as who has access to recordings and how is the recorded information accessed, etc. While it is nice to have these controls in a residence, they often do not meet the standard of a government facility set of controls. Challenges such as response time, continuous monitoring by a government resource (as opposed to a private sector company), and where the data is stored (e.g., public cloud repositories for security and camera feeds, or closed circuit) are just a few of the concerns.

Once the perimeter and overall home construction is reviewed, the precise location of the classified asset(s) must be identified so that further assessment can be made by the PhySec resource. In some cases, this may require allocating a room within the residence that has no windows and preferably not attached to the exterior walls of the residence. The room may require further construction to secure the door, door frame, internal walls, WiFi signals, listening devices or digital assistants (e.g., Amazon Alexa, Apple Siri, or Google Assistant), and more. In some cases, these types of rooms may not exist in a residence. In many cases, the use of a security container may be required to store the classified asset(s) and safeguard the injury level of the data-at-rest or access to the data, and such a security container may require a near permanent attachment to a concrete surface.

Even after noting these deficiencies, the government may not have the ability to fund the construction or procure the assets required in civil servant's residence, as there are further concerns for liability as well as differences between ownership and rental of the residence, along with the optics of spending of taxpayer monies to fund the construction in a civil servant's residence, to name just a few.

The PhySec resource will likely work closely with a government authorised PERSEC resource to ensure that all residents of the home (over a certain age) who may have access to the classified assets be cleared and granted the appropriate level of personnel clearance required (e.g., SECRET clearance). Such requirement would most likely also apply to anyone who may have access to the residence, e.g., friend, family member, maintenance company, or landlord who maintains a key and/or security alarm code to the residence.

To clear all these resources to the appropriate level required (as they may have access to the residence and therefore extended amount of time to gain access to the classified assets without being noticed) would be a difficult task under normal circumstances and take even longer during a pandemic.

Finally, the civil servant will most likely require OPSEC training to ensure they perform specific actions before use, after use, and while away from the residence hosting the classified assets. This can include where to use the assets, when to use the assets, not using them near any wireless access points, windows, into cameras, around others, disconnect other electronic equipment in the residence, and more.

Resources skilled in securing government facilities will most likely mark any residence and their occupants as an elevated risk and not recommend the location for consumption or creation of classified information, or storage of assets that may access classified resources.

As can be ascertained from the content above, PhySec, OPSEC, and PERSEC evaluations should be performed by trained and skilled individuals to measure the likelihood and thereby risk of a variety of matters including (1) the theft of the access and assets to (2) coercion by another party to the sale of access (or the assets) by the civil servant to another party.

For a government department to measure these risks, they would likely need a large team to visit the residences, interview the civil servant, and ensure mitigations are in place prior to granting secure remote access, raising the follow-up question: who is responsible to fund the travel and these mitigations? All the while the government is likely strongly advocating physical distancing policies and enforcing inter- and intra-city travel restrictions.

Under these travel restrictions, it would be challenging for an adversary to physically target the civil servant at their home (e.g. parabolic dish with microphones from inside nearby vehicles, low-powered voice and/or video capture and transmitting devices located in on windows, or walking up to the property and residence), however, if for example the civil servant was considered a high value target prior to the pandemic and resides in a high-rise or other shared complex, the neighbouring apartment may already be occupied by a well-funded adversary.

The non-technical security threats and risks that government facilities face are well understood and managed by the traditional security community. Safeguards that can be easily employed at an office are very challenging to construct and impose in a home, and in may even be restricted by state laws.

3. Technological security

Today's homes are filled with technological capabilities that would never be present in a government facility where matters of national interest are discussed and shared. Voice controlled home automation (e.g., Amazon Alexa, Apple Siri, or Google Assistant), personal security and alarm systems, poorly- or un-maintained residential hardware and software products, in-home and neighbouring Wi-Fi access points, electromagnetic interference, poor password controls and more plague the residence and consumer market.

At a high level, a possible communication path from the residence to the office could be (1) a State controlled and authorised national interest information compute device (e.g. terminal, phone, computer) communicating via (2) a state controlled and authorised cryptographic assets (algorithm, keys, and hardware) protected by (3) a state controlled and authorised firewall to (4) a government compute facility that mirrors the layers of the Communications Security (COMSEC) and IT Security (ITSEC) described.

The Figure 1 below depicts a high-level example of a classified WFH kit (e.g., SECRET) supporting information communications in red and state encrypted and safeguarded communications (via algorithms, keys, and hardware) over public networks in black communicating to-and-from the office.

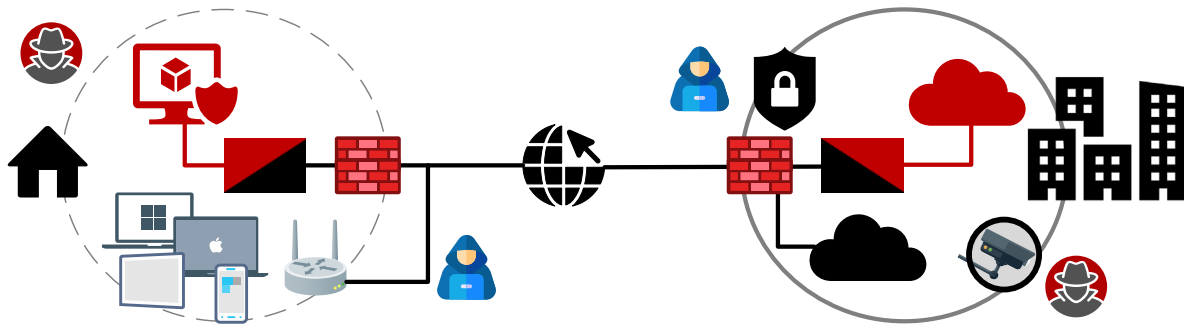


Figure 1: High level example of a secure communication path from the home to a government office

In such a case, home users would be granted access to classified networks (e.g., SECRET) using a classified WFH kit via a user account with defined security controls to and through the respective authorised classified device located in their residence. Governments may decide to grant new remote-access-only accounts created specifically with very limited access to national interest information systems and very limited functionality (e.g., no printing, no injection or extraction of data, mandatory card-and-PIN access) to minimise the possible exposure of confidentiality and adverse impact to data integrity. Governments may also impose additional ITSEC policies and safeguards using mature security logging, monitoring, analytics, alerting, active defence, multifactor authentication, zero-trust capabilities, and other technologically supported capabilities.

No matter what technological oversight (security monitoring) or controls used, certain mitigations from the home will simply not be possible without correlation of information from PhySec, and even then, the risk of information exposure could remain high. For example, it would be impossible to know if the resource *actually used* their credentials to access the system or if another person is using it on their behalf (insider threat).

3.1 Scenario 2: Insider threat

A common cybersecurity risk to any organisation is insider threat. As per recent research, almost 30% of cyber attacks are estimated to be conducted by an insider, but insider threats need not be cyber-based (see https://resources.sei.cmu.edu/asset_files/SpecialReport/2005_003_001_51946.pdf).

A civil servant using a classified WFH kit could choose to turn against their state by:

1. capturing photographs or video camera footage of the on-screen classified information, and/or
2. inviting foreign state adversaries into the residence to use the device and access and possibly even alter national interest information through the civil servant's credentials.

In either of these cases, there is little the government can do to notice this behaviour through IT security monitoring of the access by the civil servant from their residence.³ Even when using very mature methods of IT security monitoring, correlation, analytics, and forensic analysis, these two behaviours mentioned above would not be detected as they exhibit expected and normal use of a common user's access. Furthermore, regular use of search tools by the user and opening random documents becomes the baseline for this user, then behaviour analytics tools may overlook this user's behaviour. If the user is called in an interview to validate their actions, they could cover it by stating the available information management search tools are terrible and force the user to open and read too many documents to find the required information. While limits and access controls applied to user accounts may be considered good practise, it is not necessarily common practise.

3.2 Scenario 3: RF and EM interference

At a radio frequency and electromagnetic level, when commercial off the shelf (COTS) equipment is used for the end-user compute device in the residence, the electromagnetic signals emitting from the device's display (Kuhn 2005) and each keystroke (Vuagnoux and Pasini 2009) can be captured and easily deciphered by a technically inclined high-school or college student using equipment readily available from common Internet sources mixed with some home-made antennas. Depending on the type of residence, some EMSEC exploits (spatially and conducted emanations) can be particularly obvious to spot.

³ Breach of confidentiality could be to a foreign state, multi-national or domestic corporation for corporate advantage, or even anonymously to the public. The breach of confidentiality need not be for financial gain.

If TEMPEST (National Institute of Standards and Technology (NIST): Computer Security Resource Center (CSRC) n.d.) end user equipment is being considered to mitigate EMSEC attacks at the residence, the physical security of the TEMPEST tested and certified equipment becomes very important. The user of the equipment will need to remain aware of any tamper to the equipment and be vigilant in carrying out all PhySec policies and procedures along with possibly state COMSEC doctrine. If the state is unable to adequately control the physical space where the TEMPEST equipment is located and used, then the use of TEMPEST is of little use. Adversaries with close access to the TEMPEST equipment need not use EMSEC methods to gain access to the national interest information. Various low-cost screen capture and keyboard capture/injection devices can be purchased online that are easily connected inline with display and keyboard cables. These capture and injection devices can allow an adversary to record any activities, be notified when end user activity is detected, and watch live or review all the recorded activity from a repository in the cloud using commercially available tools such as “Screen Crab” and “Key Croc” (Hak5 n.d.).

3.3 Scenario 4: Use of COTS

Governments may believe COTS equipment, COTS cryptographic algorithms, and certificates are adequate for safeguarding their national interest information under the WFH policy, possibly with other safeguards and limited access. In such cases, the use of side channel attacks and fault injection attacks offer a remote method to capture cipher keys (secret keys) from compute devices when they are normally used by the residential user using commercially available tools such as the Rambus DPA and SPA workstation side channel attack platform (Rambus Inc. n.d.).

These side channel attacks use differential power analysis of a cryptographic device. As an example, devices need to be powered on, can be in “airplane mode,” and accessing their secret key. With enough data, today’s compute capacities can detect very small similarities and extract AES, RSA, DES, and other ciphers on various devices including FIPS 140-2 and 140-3 certified units. With a cipher key in hand, the adversary opens the door to other options for espionage and sabotage. Side channel attacks can be effective with various compute devices (e.g., mobile devices, dedicated cryptographic devices, computers, Internet of things, terminals).

For example, if states only consider some of the security domains (e.g., COMSEC and ITSEC) and focus on adversaries breaching their systems primarily through mathematical means, they will likely miss out on safeguarding the national interest information. When one considers the types of adversaries trying to acquire national interest information, ignoring these risks can prove to be very dangerous.

4. Discussion

During the questions and answers after his speech on August 8, 2017, at the annual *Space and Missile Defense Symposium* in Huntsville, Alabama, General John Hyten, now Vice Chairman of the Joint Chiefs of Staff and in 2017 commander of US Strategic Command (USSTRATCOM), stated (U.S. Strategic Command 2017):

“There’s no such thing as a war in cyberspace. There’s just war. We have to figure out how to defeat our adversaries, not to defeat the domains that they operate in. So we need capabilities and we need technologies to stay ahead of our adversaries in each one of those elements.”

His quote was in response to a question on the USSTRATCOM’s priorities. When applied to securing government information during WFH, defending one security domain at a time (PhySec, OPSEC, PERSEC, ITSEC, COMSEC, or EMSEC) is not a good defence. Adversaries will attempt to breach all domains at the same time.

Combining the non-technological and technological domains, adversaries can not only gain access to more information (deeper into the organisation or individual), but also use these capabilities to gain access to their target’s trusted parties (wider scope). A good example of this is the recent supply chain integrity vulnerability that we all experienced via the SolarWinds cyber attack, impacting nearly 18,000 of their customers (Porter 2021). In this case, the adversary’s target was not just SolarWinds (Tidy 2020). It was most likely many of the suppliers of government services, e.g., FireEye (FireEye Inc. 2020) and Microsoft (MSRC Team 2020), to allow for multiple points of entry into 250 branches of governments and suppliers (Porter 2021).

Under the political challenges of the coronavirus pandemic, governments around the globe have made concessions so that their governments may continue to govern, as the US Army did when they deployed access to SIPRNet for over 2000 users (Eversden 2020). The priority to (1) maintain public health and the ability to

earn/spend, (2) uphold international agreements, and (3) maintain a reliable economy with a stable state currency is critical. Losing any one of those three will crumble the remaining two.

At the February 17, 2021, UN Security Council open meeting on *Ensuring Equitable Access to COVID-19 Vaccines in Contexts Affected by Conflict and Insecurity*, UN Secretary-General António Guterres stated (Guterres 2021):

“Yet progress on vaccinations has been wildly uneven and unfair. Just 10 countries have administered 75 per cent of all COVID-19 vaccines. Meanwhile, more than 130 countries have not received a single dose.”

Secretary-Generals Guterres’ remarks are a reminder that if states do not work together to manage the coronavirus pandemic, only a small number of states will have administered the COVID-19 vaccine broadly (10 countries). That means, many states will not be positioned to retract their “national interest WFH” decisions for a long time.

For example, the costs to retrofit a government building to mitigate risks of all security domains is very high, and the timeline to achieve these safeguards is significantly longer than deploying a classified WFH kit to under 100 WFH civil servants. Indeed, they are not the same, but in the latter case, government operations can immediately deliver value to its citizens, whereas in the other, it could take years to begin. As states release a classified WFH capability, much like the US Government has done with SIPRNet, they may wish they never opened Pandora’s box.

In October 2019, the Global Health Security Index report (Global Health Security Index 2019) measured 195 countries for their ability to survive a global epidemic or pandemic – based on their preparedness and playbooks. The actual results thus far during the coronavirus pandemic differs greatly from that report’s projections. Have such preparations and playbooks provided false guidance to governments? We can only hope the pandemic serves as a learning opportunity for governments to review federal disaster plans, test their capabilities, and spend the time to pre-think for the future.

5. Conclusion

We live in a world where certain governments have accepted the high risks to a variety of threats with their WFH policies. If a foreign state sponsored adversary makes their way in, who knows how long they will be able to linger and how deep and wide their impact will be. Maybe, by the time any breach surfaces, democratic governments will have changed hands. Regardless of how the governments react and respond after the breach, the impact to confidentiality, integrity, and access may have already been realised.

The global coronavirus pandemic has mostly resulted in a chaotic response from governments and corporations with their respective WFH policies. These WFH methods may seem adequate for matters of non-national interest, however, even with encryption safeguards, national interest WFH has led us to wonder “what the heck” have we done.

States whose citizens are near fully vaccinated can return to processing matters of national interest in the safety and safeguards offered by their government facilities, hopefully terminating any classified WFH capabilities that may have been adopted. The longer it takes for states to vaccinate their citizens, the longer the government national interest WFH policies will remain in effect, leaving (1) governments as an even bigger target for domestic and foreign espionage and sabotage, and (2) the civil servants with classified WFH access an obvious and arguably easier target. Although not discussed in this paper, similar challenges can be extended to critical national infrastructure and industries.

References

- Canadian Centre for Cyber Security. 2018. *IT Security Risk Management: A Lifecycle Approach (ITSG-33)*. 05 November. Accessed December 12, 2021. <https://cyber.gc.ca/en/guidance/it-security-risk-management-lifecycle-approach-itsg-33>.
- Carvin, Stephanie. 2021. “Canadian National Security Operations during COVID-19.” Chap. 6 in *Stress Tested: The COVID-19 Pandemic and Canadian National Security*, 107-126. Calgary: University of Calgary Press. Accessed December 12, 2021. https://prism.ucalgary.ca/bitstream/handle/1880/114134/9781773852447_OA.pdf?sequence=1#page=117.
- Estonian National Security Authority. n.d. *Comparative Tables*. Accessed December 12, 2021. <https://www.valisluureamet.ee/nsa/tables.html>.

- Eversden, Andrew. 2020. "The Army will soon allow users to access classified info from home." *C4ISRNet*. 22 June. Accessed December 12, 2021. <https://www.c4isrnet.com/2020/06/22/the-army-will-soon-allow-users-to-access-classified-info-from-home/>.
- FireEye Inc. 2020. "Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor." *Mandiant Inc.* 13 December. Accessed December 12, 2021. <https://www.mandiant.com/resources/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor>.
- Giri, Susma. 2022. "The Impact of Covid-19 on National Security of Nepal." *Unity Journal* 2: 236-264. Accessed December 12, 2021. doi:<https://doi.org/10.3126/unityj.v2i0.38852>.
- Global Health Security Index. 2019. "2019 Global Health Security Index." 20-29. Accessed December 12, 2021. <https://www.ghsindex.org/wp-content/uploads/2020/04/2019-Global-Health-Security-Index.pdf>.
- Guterres, António. 2021. "Secretary-General's remarks to the Security Council Open Meeting on Ensuring Equitable Access to COVID-19 Vaccines in Contexts Affected by Conflict and Insecurity [as delivered]." *United Nations*. 17 February. Accessed December 12, 2021. <https://www.un.org/sg/en/content/sg/statement/2021-02-17/secretary-generals-remarks-the-security-council-open-meeting-ensuring-equitable-access-covid-19-vaccines-contexts-affected-conflict-and-insecurity-delivered>.
- Hak5. n.d. *Hak5 Official Site*. Accessed December 12, 2021. <https://shop.hak5.org/>.
- Kuhn, Markus G. 2005. "Electromagnetic Eavesdropping Risks of Flat-Panel Displays." In *Privacy Enhancing Technologies: 4th International Workshop, PET 2004, Toronto, Canada, May 26-28, 2004. Revised Selected Papers*, 88-107. Berlin: Springer. Accessed December 12, 2021. doi:https://doi.org/10.1007/11423409_7.
- MSRC Team. 2020. "Microsoft Internal Solorigate Investigation Update." *Microsoft Security Response Center (MSRC)*. 31 December. Accessed December 12, 2021. <https://msrc-blog.microsoft.com/2020/12/31/microsoft-internal-solorigate-investigation-update/>.
- National Institute of Standards and Technology (NIST): Computer Security Resource Center (CSRC). n.d. *Glossary: TEMPEST*. Accessed December 12, 2021. <https://csrc.nist.gov/glossary/term/TEMPEST>.
- Porter, John. 2021. "White House now says 100 companies hit by SolarWinds hack, but more may be impacted." *The Verge*. 18 February. Accessed December 12, 2021. <https://www.theverge.com/2021/2/18/22288961/solarwinds-hack-100-companies-9-federal-agencies>.
- Rambus Inc. n.d. *DPA Workstation Analysis Platform*. Accessed December 12, 2021. <https://www.rambus.com/security/dpa-countermeasures/dpa-workstation-platform/>.
- Tidy, Joe. 2020. "SolarWinds Orion: More US government agencies hacked." *BBC News*. 15 December. Accessed December 12, 2021. <https://www.bbc.com/news/technology-55318815>.
- Treasury Board of Canada Secretariat. 2019a. *Directive on Security Management - Appendix J: Standard on Security Categorization*. 01 July. Accessed December 12, 2021. <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32614>.
- . 2019b. *Policy on Government Security, Chapter 5*. 01 July. Accessed December 12, 2021. <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16578#cha5>.
- U.S. Strategic Command. 2017. *General John E. Hyten Speech, Space and Missile Defense Symposium*. 08 August. Accessed December 12, 2021. <https://www.stratcom.mil/Media/Speeches/Article/1274339/spaceand-missile-defense-symposium/>.
- Vériter, Sophie L., Monica Kaminska, Dennis Broeders, and Koops Koops, . 2021. *Responding to the COVID-19 'infodemic': National countermeasures to information influence in Europe*. The Hague: The Hague Program for Cyber Norms. Accessed December 12, 2021. <https://www.thehaguecybernorns.nl/news-and-events-posts/responding-to-the-covid-19-infodemic-national-countermeasures-against-information-influence-in-europe>.
- Vuagnoux, Martin, and Sylvain Pasini. 2009. "Compromising electromagnetic emanations of wired and wireless keyboards." *18th USENIX Security Symposium*. 1-16. Accessed December 12, 2021. https://www.usenix.org/legacy/events/sec09/tech/full_papers/vuagnoux.pdf.

The Failure of Trust in Trusted Systems

Angela Mison, Gareth Davies and Peter Eden

University of South Wales, Treforest, UK

Angela.Mison@southwales.ac.uk

Gareth.Davies@southwales.ac.uk

Peter.Eden@southwales.ac.uk

Abstract: Contrary to the early days in which freedom of access and knowledge for all was a fundamental tenet of the burgeoning internet, today, the internet is a hostile environment. This paper represents that practices of trusted autonomous systems and zero trust must reside both organisationally and in any connected device, and that they will be insufficient to secure any hyperconnected system. For example, in the area of Connected and Autonomous Vehicles (CAV), the emphasis of standards associated with cybersecurity is to protect vehicles from an external attack, particularly where it may have an impact on safety. The Internet of Vehicles (IoV) is a subset of the IoT. Security of the IoT has not been standardised and is applied by proprietary organisations. Little consideration has been given to the IoV not being the target of an attack, but a means to an end. Vehicles necessarily communicate with the infrastructure and other vehicles. Additionally, outsourced organisational systems and third-party components within a complex, interconnected, communicating system renders it impossible to define and secure all endpoints. Malware in any part of the hyperconnected systems, provides opportunities for hackers. The security of the systems is only as strong as its weakest link. As an example of an onward attack, a DDOS attack is a debilitating exhaustion of resources attack that disables intended operation of a system. Formation of a botnet from the IoV for use in onward attacks is hypothesised. Such a botnet could have a global reach. Research has indicated that the complete cleansing of a computer-based botnet could take between 5 and 15 years. However, with decentralised command and control and peer to peer communication, the botnet could remain persistent. It is shown that persistent botnet formation software is readily available on the dark market and that specialist software can be commissioned as Crime as a Service. Organised crime groups have already reverse engineered vehicle systems. Mobile attack platform swarms are an attractive proposition to the malefactor, and with the potential for global reach, perhaps the next step on from ransomware. Intermittent disruption from mobile sources is hard to trace and there is an existing pool of 1.2 billion vehicles.

Keywords: IoV, Trusted autonomous systems, zero trust, botnet, DDOS attack platform, cybersecurity

1. Introduction

Contrary to the early days in which freedom of access and knowledge for all was a fundamental tenet of the burgeoning internet, today, the internet is a hostile environment. Hyperconnectivity, resulting from the formation of a system of systems, has introduced major security difficulties, some of which are not covered by trusted autonomous systems or zero trust mechanisms. The example used throughout this paper is that of a vehicle, a device in the Internet of Vehicles (IoV) which itself is a subset of the IoT and also possesses industrial control system attributes (Stouffer, et al. 2015).

Vehicles have intelligent interconnected systems. They are subject to frequent updates by the manufacturer. It is not known, contractually, if third-party components can be updated by their suppliers. Overall, they represent a complex system which has been developed in an evolutionary manner in an attempt to make vehicles safer. It is not necessarily the vehicles that need to be made safer as they are responsible for only 5% of accidents. The systems interconnect and work to mitigate the actions of drivers who are inherently unsafe and directly responsible for 95% of accidents (Office for National Statistics 2021), (Transport Scotland Commdhail Alba 2016), (Rollison, et al. 2018). Each introduction of mitigation increases the complexity and can increase the vulnerability of the overall system.

When considering cyber-attacks and their defence, the emphasis is, inevitably, on defence of the device under consideration and what damage can be done as a result. While system developers think in terms of what a system should do in terms of its requirements specification, hackers think in terms of whether access to one system provides access to another and what those systems can be made to do after access has been gained.

The standards associated with cybersecurity of vehicle systems are designed to protect vehicles from malefactors and to ensure continuing safety (ISO/SAE 2021). Further discussion covers distributed responsibilities and damage scenarios which result in adverse consequences in the context of the vehicle or other road user. The latter is subject to some debate regarding the roles and responsibilities for cybersecurity

and safety. This debate provides a gap through which some considerations may fall. The industry appears to be taking the risk mitigating approach of cyber resiliency engineering (ISO/IEC/IEEE 15288:2015, Systems and software engineering—Systems life cycle processes 2015), (Ross, et al. 2021) for the vehicle itself without considering its potential for weaponisation (Horowitz 2021).

Analysis has been carried out on the potential targets and classes of hacks that might occur, but these are vehicle centric (Ghosh, 2016), With technological advances, other potential attacks have been identified (Guardknox 2021), but again these are vehicle centric. Hackers and criminals gain entry then consider monetisation opportunities (Rajput 2018).

2. Why onward attacks are feasible

Log4j (Mitre Corp e; 2021) has highlighted that zero day vulnerabilities will always represent a weakness in large interconnected systems and prove the maxim that testing shows the presence of errors, never their absence. This vulnerability demonstrated that not all historically developed systems were/are capable of being identified and updated without extensive delay. The view taken here is that the initial failing of cybersecurity may well occur in the manufacturer's system or other permitted communicating system which has been infiltrated by malware and that as a trusted source of coding and updating for its marque or application, has transmitted relevant malware to a vehicle. The vehicle itself has not been subject to attack although this may be done through firmware vulnerabilities (Regenscheid 2018) or Supply Chain vulnerabilities (Boyens, et al. 2021).

From the perspective of a hacker, it is simple to determine that vehicle systems and the context in which they operate are ill-defined, hence the existence of vehicle to everything (V2X) communications' implication that there is an onward attack route and manufacturers may be relying on the filtering security of the receiving system for communications from a vehicle. Spreading throughout the vehicle population outside specific marque and model is hypothetically possible via universally permissible vehicle to vehicle (V2V) communication.

Consider the Internet of Vehicles (IoV) simply as computers on wheels, connected to the internet via V2X. This could enable an almost undetectable source from which to launch an attack as, being mobile, it is likely that any attack would be intermittent, the hardest form of fault to find. It is not necessary to interfere with the vehicles' actual operations, but vehicles could be made to run interference against other external computer operations. The vehicle would still conform to the relevant cybersecurity standards (ISO/SAE 2021) as the vehicle itself is not under attack. Many attacks are possible, as has been demonstrated in the IoT. Lack of consideration of the IoV being a subset of the IoT may have led to complacency as to targets (Heubl 2019), (Matthews 2018) and motivation of malefactors .

3. Legal liability for onward attacks

Using the argument of the IoV as a subset of the IoT, there is potential legal liability for vehicle manufacturers as the US Federal Trade Commission (FTC) report concerning a hack into IoT devices argued that responsibility for security of a device is the responsibility of the manufacturer of the device rather than the owner/user which runs counter to the (ISO/SAE 2021) concept of distributed responsibility. The FTC filed a complaint alleging that an IoT device vendor:

failed to take reasonable steps to protect their [product] from widely known and reasonably foreseeable risks of unauthorized access (US Federal Trade Commission, 2017)

In the settlement of the case (US Federal Trade Commission, 2019) the FTC included the statement:

the settlement includes protections for consumers who currently own devices covered by the order. [The manufacturer] must automatically push fixes to devices set up to receive them and must provide clear step-by-step instructions to all consumers explaining how to patch their devices themselves.

The FTC rulings mark a distinct difference from vehicle manufacturers considering updates to vehicle systems in which there is an attempt to shift liability for non-performance, to the owner/user of the vehicle. There is industry evidence that users (owners/drivers) will prioritise updates for a useful application over a security update. An example was given of response rates to a safety system update versus an automatic parking app

offering. The take up of the parking app was much greater than the safety update, which was deferred. The argument of safety and security -v- convenience, particularly if that safety and security is not directed at the owner/user of a vehicle but to some potential remote and unknown third party, prioritises convenience – a selfish perspective (the-autonomous.com and Infineon 2020). Such procrastination is a *widely known and foreseeable risk* as defined by the FTC and should represent a major concern for manufacturers.

Also, in (the-autonomous.com and Infineon 2020) it was suggested that some parts of the system might not need as much defensive preplanning as others. There is a fundamental flaw in this argument. The primary concerns are that the overarching system itself is not defined, and the longevity of a vehicle. It has been suggested that the lifetime of a car is approximately 20 – 30 years and that vehicle identification numbers (VIN) used for official registration, and manufacturing numbers are reserved for up to 60 years before reuse (National Archives 2021).

4. The need for Fortified Cyber Defence – endpoint protection

The IoT is categorised as a set of devices that have embedded systems that connect to the internet to extend their functionality and to allow firmware to be updated and managed remotely. In a sense it is a class of pervasive computing which includes sensors, communication devices and control mechanisms that can collect data, communicate with users and be remotely controlled (Voas, Laplante and Applebaum 2018). IoT security is considered in (Fagan, et al. 2021)

However notionally complex, a vehicle is a mobile IoT device, which by its very nature is connected to the road transport system infrastructure, other vehicles and the inimical V2X. Vulnerabilities arise from the manufacturer's hyperconnectivity, through the supply chain, Cloud, any app, or any system to which it may be permissibly connected. The lack of current definition means that there is no possibility of enumerating and defending the potentially infinite number of endpoints available for the initial attack. Vehicle access points are a subset of endpoints, whether defined, within the manufacturer's and road transport system's hyperconnected system of which the vehicle is a part. It is not possible to secure and defend something which has no context or system boundaries (Burkacky, et al. 2020).

Discussions of the motivation of malefactors are mainly directed at hacking vehicles, and the reason given for any malware enabling DDOS attacks within the vehicle is given as lack of authentication (SAE International, 2017). As stated previously, within existing standards, opportunity, and liability for attacks in which vehicles are formed into a botnet for onward attacks, e.g. DDOS attack platforms, are not considered.

Drawing from Secure Beyond Breach (Reiber, 2020) and (Burkacky, et al. 2020), the first line of cyber defence is a firewall which, does not present a challenge to the skilled hacker. What is required is hierarchical fortified cyber defence (Dewar 2014). While multifactor authentication may be possible, each level of fortified defence places an additional economic burden on the hacker. The malefactor is having to breach each effective firewall. Unfortunately, each breach permits an assessment, reconnaissance of the internals, the discovery of the networks and systems, enabling the decision regarding the usefulness for offensive e.g. payload delivery, or defensive purposes e.g. kill processes, obfuscation.

5. The need for zero trust – Microsegmentation and API security

Currently, no one programs or builds a vehicle and its systems from scratch. Various parts of the systems and componentry are outsourced. Effectively, it has become a standard supply chain. The componentry is tested, validated, and certified as conforming to specification. The manufacturer has to take that on trust. It is well known that in vehicles comprising millions of lines of code, and third-party components, there will be errors, no matter how carefully and proficiently put together. The manufacturer is dependent on suppliers and their quality systems. Certainly, there are few outright failures, but opportunities for hackers can arise.

Microsegmentation involves the definition of virtual networks according to necessary communication. Any internal virtual network is a microcosm of the complex system. Prevention of incursion by an interloper does not entail consideration of benign or malignant intentions. The environment and operating parameters must be considered in as comprehensive a way as possible (Rose, Mitchell and Connelly 2020). The cybersecurity issues concerning vehicle manufacturers are the identification of interconnections which must be permitted, and the impossibility of defining the potentially infinite endpoints, some of which have been revealed to be in

the componentry (JSOF 2020). The system is not closed with defined entry and exit points. There are concerted efforts to prevent unauthorised access to systems, but having gained access, it can be assumed that hackers have the necessary authority to make external communications via undefined V2X or V2V.

Security at the API, authentication, can be achieved by having operational contours which define acceptable parameter lists, both for functional entry and exit, and which on exit are checked as input with the receiving API. This is still insufficient. There is an argument for random factors to be added to the parameter list which are difficult for a hacker to guess or calculate, but a proposed solution is the introduction of immutable systems with extended parameter list to include pattern of life factors. The implication of this is lack of access to system componentry by third parties, and removal of any on board learning facility, as any change would involve failure to match the pattern of life characteristics. The pattern of life factors could be based on timings, latency, frequency, length of the parameter list, etc. It would render the extended parameter list impossible to guess, and the rate of system updates would limit the temporal window available to hackers.

It is essential to understand that anything done in the cybersecurity and authentication field is at best a deterrent to an incursion, not a guarantee of impenetrability. Its purpose has the benefit of buying time for the defender, during which time, it is hoped that an attack, or precursor, will be detected/noticed/mitigated, and the attacker deterred from onward action. In a real time system, in a vehicle, which has limited onboard resource, incursion may not be recognised until the vehicle fails or its telematics data indicate to the manufacturer that there is a problem. Any active malware may present as a drain on resources. Detection of a problem in the system does not indicate the location of the problem, merely its consequence.

6. Botnets/Thingbots

A botnet is a logical collection of connected devices in which command and control of devices can be exercised by an unauthorised third party. A thingbot, though a term not in very common usage, refers to a botnet of compromised IoT devices. (IBM X-Force® Research, 2017). A botnet is designed to carry out the tasks ordered by the botmaster. By enabling communication from a vehicle, it is possible to use the fleet as a platform for a DDOS attack or any other attack.

With the number of vehicles on the road, the degree of frequent communications between V2X and V2V, and the price of botnets being dependent on their quality and number of infected devices, the arrival of the suggested attack platform is late in coming. This tardiness may be reflective of the protections applied and the complexity of the architecture, given that there is other low hanging fruit in the IoT environment. It may already be present, in the fleet, and if so, its persistence is detrimental.

Botnets are common in criminal usage. There are many botnets for lease in the Dark Market. They are used to commit crimes using someone else's computing or other resources and, depending on how they are encoded, can make it difficult to trace back to the perpetrator. The malware to construct a botnet, including demonstration videos of the malware in action can be found on youtube.com. There are sites in the dark web such as AlphaBay, reputedly closed in 2017 (Dark Web Link, 2019), which approximate to eBay for criminals, where it is possible to purchase the malware of choice for any particular exploit, or to commission same.

A thingbot of compromised vehicles can themselves form the basis of an onward attack, in this instance DDOS is the exemplar. There is a reasonably standard process for creating and growing botnets. Increasingly, Peer to Peer (P2P) protocols are used (Park, Ratzin and van der Schaar, 2011). P2P is a networking mechanism that requires no central server. Multiple links are available between all members of the network. A P2P system would compare version numbers of the botnet and update to the newest version between connected peers leading to the botnet being in a constant state of renewal. The result is resilient and pervasive.

7. Botnet establishment example: dark_nexus

Bitdefender have authored a white paper (Bitdefender, 2020) concerning a new IoT botnet, dark_nexus, influenced by Qbot and Mirai with original core modules. dark_nexus comes from a known malware author and seller, greek.Helios and has been subject to approximately 30 updates in 3 months, dating from December 2019 and March 2020, enhancing its capabilities and robustness.

Although greek.Helios does not appear to offer full Crime as a Service openly, dark_nexus is available for lease on a Dark Market equivalent to eBay from \$20 upwards starting from 2500 seconds (Dark Web Link, 2019). One of the uses of dark_nexus has been DDOS attacks. To date most botnet victims would appear to be in the Far East, Eastern Europe, and Brazil, but with IoT, spread can be global as represented in the Akamai quarterly report (Akamai, 2020), (Akamai a, 2020), (Arghire, 2020).

Hash values for each known version are available, on GitHub as Indicators of Compromise. The rate of versioning would seem to indicate that while older versions may be detected automatically, other detection mechanisms must be used in addition.

Thus far, dark_nexus propagates by brute forcing Telnet protocol, an economically profitable time and effort method for breaching susceptible systems. The botnet has been customised for 12 architectures and its delivery is dynamically configured for the victim system. Inter alia it introduces code enabling SOCKS5 proxy (Leech, et al., 1996) which operates at Layer 5 (Session) over TCP/IP that enables tunnelling and supports IPv6. dark_nexus botnets operate using asymmetric encryption for communications between the device and the C2 server.

dark_nexus incorporates a risk assessment process that identifies other processes which may represent a threat to its existence. It maintains a whitelist of processes and their IDs and terminates processes that cross a threshold value of risk it has calculated. It is thought to do this in order to ensure monopoly control over its botnets, thereby ensuring price stability and control for its author and users.

The development of such software may be a skilled artform, and take much time, but there is no reason to suspect that the economic return on investment by the malware author will not repay any investment of time spent researching the possibilities.

8. Supply chain introduced vulnerability: Ripple20

To emphasise the potential problems associated with the use of third-party components, and supply chain similarities, the example given is that of Ripple20. This extends the concept beyond the IoV and into the national critical infrastructure.

Ripple20 (JSOF, 2020) presents nineteen zero day vulnerabilities in a widely embedded TRECK TCP/IP stack. There have been historical upgrades to the stack, but with the number and age of systems in use, not all the affected stacks will have been patched or upgraded, and differing configurations and collaborations add to the complexity of the issue.

The vulnerabilities identified enable attackers to bypass NAT and firewalls. They permit the undetected control of devices without user interaction. Any communication packets that may result are similar to or are actual valid packets. Resultant traffic may thus appear legitimate. JSOF has demonstrated exploitation of these vulnerabilities on different devices as a proof-of-concept. (JSOF, 2020)

Fifteen of the vulnerabilities have ratings scores ranging between 3.1 and 8.2. Four of the most serious vulnerabilities have a CVSS score >9 and are indicated below:

- CVE-2020-11896 enables remote code execution related to IPv4 tunnelling (Mitre Corp a, 2020).
- CVW-2020-11897 The Treck TCP/IP stack before 5.0.1.35 has an Out-of-Bounds Write via multiple malformed IPv6 packets (Mitre Corp b, 2020)
- CVE-2020-11898 The Treck TCP/IP stack before 6.0.1.66 improperly handles an IPv4/ICMPv4 Length Parameter Inconsistency, which might allow remote attackers to trigger an information leak (Mitre Corp c, 2020)
- CVE-2020-11901 The Treck TCP/IP stack before 6.0.1.66 allows Remote Code execution via a single invalid DNS response (Mitre Corp d, 2020)

A more detailed indication of the vulnerabilities is given in (JSOF, 2020) from which it can be deduced that these vulnerabilities may represent a critical threat to vehicles, resulting in their use as attack platforms, both onward and internally. Some useful mitigation advice is given.

Design reliance on a specific embedded system may render upgrades infeasible, particularly where legacy systems may be involved. There are reports indicating that a full system update can take 5 – 15 years across affected devices before the total population is treated. Many of the systems affected by Ripple20 do not directly contain TRECK TCP/IP stacks, and may not even be connected to the internet, but any resultant infection may have been imported through, or amplified by, the supply chain. (JSOF, 2020).

For an embedded system which is highly defective, widely distributed, highly vulnerable and has been in existence for in excess of twenty years, it is very unlikely that all devices, quoted as in the billions, will or can be updated, which leaves the IoT and the national critical infrastructure with a persistent and potentially pervasive problem, whereby an unknown population of raw devices will not or cannot get the necessary firmware update.

9. Maintenance ramifications and recovery from attack

Although no one can state with any certainty the types and numbers of attacks that may become possible during a possible 60 year lifetime, it is certain that with technological developments some, at least, will become foreseeable. Timeliness of security updates becomes a critical maintenance issue as it may not be possible to update during vehicle operation without disabling the vehicle or leaving the embedded systems and neural networks in consequential unknown states. The problem is exacerbated by differentiated timing of updates leaving an apparent single class of vehicles in a mixed, unknown system state which could lead to incompatibility and inconsistency. A leaf could be taken from the hackers and updates effected through use of a botnet, a copy of the new immutable system, determined by version number, being made to an additional hard drive for update at a convenient time.

If it cannot be demonstrated that manufacturers have developed their systems with due care and consideration of regulations, standards, safety, security, and reasonably foreseeable events, they are likely to fall foul of their insurers and open themselves to class actions. Of particular relevance to manufacturers will be the semantics of the verbs in the standards. The standard SAE J3216 (SAE International, 2020) seeks to provide a common semantics for the industry so that to promote a common understanding. Where a recommended standard's requirement has not been filled, it will be for the courts to decide the reasonableness of that non-performance.

10. Conclusion

Although this paper addresses the potential for onward attacks in the IoV, that potential exists in all IoT systems. With promotion of evermore interconnectedness, the time may have come to consider the way the internet is seen and used. Considered definition of contexts in which the IoV and critical national infrastructure work would enable microsegmentation, the definition of end points, and the ring fencing of operational platforms. Ring fenced, the endpoints are not infinite, but defined by the purpose.

There is a reliance on cyber defence regarding an attack against the vehicle itself, but little consideration to its use as an attack platform. One of the purposes of security by design is to consider what a system can be made to do, not what it was designed to do. Even in this, the perspective tends to the internal. There is never a guarantee in the face of a determined attacker, or against cross infection from a permissible or trusted source.

Security standards lag the development and sale of IoT devices. Existing devices may be infected and persistently so. Both owners and manufacturers may be unaware of the types of infection indicated above. Enthusiasts may exacerbate problems through modifications to their vehicles.

There are techniques which mitigate against such onward attacks but, due to persistence, the opportunities for such may continue in a residual population leading to re-establishment of the botnet. Peer to peer communication is enabled via Vehicle to Vehicle (V2V) communication, essential for collision avoidance and situation awareness, thus it is not amenable to a simple solution. It is not certain at this time how a fleet could be tested and cleansed without simultaneous update of all vehicle systems in existence. Mobility adds to the problems as even if mandated updates are undertaken in one country, crossing a border in a vehicle could reinfect the fleet.

Ripple20 has demonstrated that the supply chain represents a vulnerability to vehicles. Stringent security by design methods should be in place in the manufacture or development of any systems to be deployed in a vehicle and any resulting system approved for release subjected to a security audit. This will not be proof of an inviolate system but should be capable of identification of many vulnerabilities. It is possible that this could be enforced by SLAs and a sharing of potential liability in contractual clauses.

Any resolution needs to consider the Internet of Vehicles, a subset of the IoT, as a botnet, or launch platform for onward attacks. Security by design is a current paradigm. In that process, extended by (Reiber, 2020) there needs to be consideration of what can be done as a launch platform, how to prevent the installation of necessary malware, and how to limit V2X communications. The resolution developed should be generically applicable to all endpoints and capable of scaling to the projected 2.3 Billion vehicles on the road.

Systems evolve, bringing with them their history, and often combining operational and monetisation platforms as a post hoc activity. For vehicles, the monetisation platform profits support the original manufacturing company, and has a knock on effect within third party's asset and fleet management. Perhaps the underlying problem of cybersecurity within the IoT has resulted from growth of surveillance capitalism which has led to platform monetisation and data as a commodity.

References

- Akamai a, 2020. *Your Guide to DDOS Attacks*. [Online] Available at: <https://www.akamai.com/uk/en/resources/ddos-attacks.jsp> [Accessed 23 June 2020].
- Akamai, 2020. *Q4 STATE OF THE INTERNET SECURITY SPOTLIGHT*. [Online] Available at: <https://www.akamai.com/uk/en/multimedia/documents/social/q4-state-of-the-internet-security-spotlight-iot-rise-of-300-gbp-ddos-attacks.pdf> [Accessed 23 June 2020].
- Arghire, I., 2020. *Security Week: Akamai, Amazon Mitigate Massive DDoS Attacks*. [Online] Available at: <https://www.securityweek.com/akamai-amazon-mitigate-massive-ddos-attacks> [Accessed 23 June 2020].
- Bitdefender, 2020. *New dark_nexus IoT botnet puts others to shame*. [Online] Available at: <https://www.bitdefender.com/files/News/CaseStudies/study/319/Bitdefender-PR-Whitepaper-DarkNexus-creat4349-en-EN-interactive.pdf> [Accessed 20 June 2020].
- Boyens, J. et al., 2021. *Draft (2nd) NIST Special Publication 800-161 Revision 1: Cybersecurity Supply Chain Risk Management Practices for Systems and Organisations*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1-draft2.pdf> [Accessed 20 December 2021].
- Burkacky, O. et al., 2020. *Cybersecurity in automotive: mastering the Challenge*. [Online] Available at: <https://www.mckinsey.com/~media/mckinsey/industries/automotive%20and%20assembly/our%20insights/cybersecurity%20in%20automotive%20mastering%20the%20challenge/cybersecurity-in-automotive-mastering-the-challenge.pdf> [Accessed 20 December 2021].
- Dark Web Link, 2019. *AlphaBay Dark Web Link*. [Online] Available at: <https://darkweblink.com/alphabay/> [Accessed 24 June 2020].
- Dewar, R., 2014. *The Triptych of Cyber Security: A Classification of Active Cyber Defence..* Talinn, NATO OCD COE Publications.
- Fagan, M. et al., 2021. *NIST Special Publication 800-213: IoT Device Cybersecurity Guidance for the Federal Government: Establishing IoT Device Cybersecurity Requirements*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-213.pdf> [Accessed 20 December 2021].
- Ghosh, S., 2016. *Automotive Cybersecurity: From perceived threat to stark reality*, Warrendale: SAE International.
- Guardknox, 2021. *Vehicle Hacking is here to stay*. [Online] Available at: <https://www.guardknox.com/vehicle-hacking-connected-vehicles/> [Accessed 20 December 2021].
- Heubl, B., 2019. *How to hack an IoT device*. [Online] Available at: <https://eandt.theiet.org/content/articles/2019/06/how-to-hack-an-iot-device/> [Accessed 20 December 2021].
- Horowitz, M., 2021. *Cybercrime predictions for 2022: Deepfakes, cryptocurrencies and misinformation*. [Online] Available at: <https://www.weforum.org/agenda/2021/11/2022-cybercrime-predictions-checkpoint/> [Accessed 20 December 2021].
- IBM X-Force® Research, 2016. *The inside story on botnets*. [Online] Available at: <https://www.ibm.com/downloads/cas/V3YJYVZX> [Accessed 23 June 2020].
- IBM X-Force® Research, 2017. *The weaponisation of IoT Devices: The rise of the thingbot*. [Online] Available at: <https://www.ibm.com/downloads/cas/6MLEALKV> [Accessed 2020 June 20].
- IEEE, 2015. *ISO/IEC/IEEE 15288:2015, Systems and software engineering—Systems life cycle processes*. [Online] Available at: <https://ieeexplore.ieee.org/document/7106435> [Accessed 20 December 2021].
- Islam, M., Nguyen, G., Pan, R. & Rajan, H., 2019. *A comprehensive Study on Deep Learning Bug Characteristics*. s.l., ACM, pp. 510-520.
- Islam, M., Pan, R., Nguyen, G. & H, R., 2020. *Repairing Deep Neural Network:: Fix Patterns and Challenges*. Seoul, ACM.

- ISO/SAE, 2021. *ISO/SAE 21434:2021 (en): Road vehicles Cybersecurity engineering*. [Online] Available at: <https://www.iso.org/obp/ui/#iso:std:iso-sae:21434:ed-1:v1:en> [Accessed 20 December 2021].
- JSOF, 2020. *Ripple20*. [Online] Available at: <https://www.jsf-tech.com/ripple20/> [Accessed 20 June 2020].
- Leech, M. et al., 1996. *SOCKS Protocol Version 5*. [Online] Available at: <https://www.ietf.org/rfc/rfc1928.txt> [Accessed 4 July 2020].
- Matthews, K., 2018. *Evolution of IoT hacking*. [Online] Available at: <https://www.iotforall.com/evolution-iot-hacking> [Accessed 20 December 2021].
- Mitre Corp a, 2020. *CVE-2020-11896*. [Online] Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-11896> [Accessed 20 June 2020].
- Mitre Corp b, 2020. *CVE-2020-11897*. [Online] Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-11897> [Accessed 20 June 2020].
- Mitre Corp c, 2020. *CVE-2020-11898*. [Online] Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-11898> [Accessed 20 June 2020].
- Mitre Corp d, 2020. *CVE-2020-11901*. [Online] Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-11901> [Accessed 20 June 2020].
- Mitre Corp e, 2021. *Search Result log4j*. [Online] Available at: <https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=log4j> [Accessed 20 December 2021].
- National Archives, 2021. *Part 565 - Vehicle Identification Number (VIN) Requirements*. [Online] Available at: <https://www.ecfr.gov/current/title-49/subtitle-B/chapter-V/part-565> [Accessed 2021 December 2021].
- Office for National Statistics, 2021. *Reported road accidents, vehicles and casualties tables for Great Britain*. [Online] Available at: <https://www.gov.uk/government/statistical-data-sets/reported-road-accidents-vehicles-and-casualties-tables-for-great-britain#contents> [Accessed 20 December 2021].
- Park, H., Ratzin, R. & van der Schaar, M., 2011. *Peer to Peer Networks - Protocols, cooperation and Competition*. [Online] Available at: https://www.vanderschaar-lab.com/papers/chapter_P2P_hpark.pdf [Accessed 20 December 2021].
- Rajput, B., 2018. Understanding Modus Operandi of the Cyber Economic Crime from People-Process-Technology Frameworks Perspective. *Journal of Emerging Technologies and Innovative Research*, March.5(3).
- Regenscheid, A., 2018. *NIST Special Publications 800-193: Platform Firmware Resiliency Guidelines*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-193.pdf> [Accessed 20 December 2021].
- Rollison, J., Regev, S., Moutari, S. & Feeney, A., 2018. What are the factors that contribute to road accidents? An assessment of Law Enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis and Prevention*, Volume 115, pp. 11-24.
- Rose, S., Mitchell, S. & Connelly, S., 2020. *NIST Special Publication 800-207 Zero Trust Architecture*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf> [Accessed 2021 December 2021].
- Ross, R. et al., 2021. *Developing Cyber-Resilient Systems: A systems Security Engineering Approach*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v2r1.pdf> [Accessed 20 December 2021].
- SAE International, 2017. *CERTIFICATES AND SIGNATURES: How to Ensure Authentication in the IoT*. [Online] Available at: <https://saemobilus.sae.org/content/TBMG-26489/> [Accessed 24 June 2020].
- SAE International, 2020. *SAE J3216 Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles*, Warrendale: SAE International.
- Stouffer, K. et al., 2015. *NIST Special Publication 800-82 Revision 2 Guide to Industrial Control Systems (ICS) Security*. [Online] Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r2.pdf> [Accessed 20 December 2021].
- the-autonomous.com and Infineon, 2020. *Virtual Meeting Chapter Event Safety and Security*, s.l.: s.n.
- Transport Scotland Commdhail Alba, 2016. *Article 2 Contributory factors to reported road accidents*. [Online] Available at: <https://www.transport.gov.scot/publication/reported-road-casualties-scotland-2016/article-2-contributory-factors-to-reported-road-accidents/> [Accessed 29 December 2021].
- US Federal Trade Commission, 2017. *FTC sues D-Link over router and camera security flaws*. [Online] Available at: <https://www.consumer.ftc.gov/blog/2017/01/ftc-sues-d-link-over-router-and-camera-security-flaws> [Accessed 20 June 2020].
- US Federal Trade Commission, 2019. *D-Link settlement: Internet of Things depends on secure software development*. [Online] Available at: <https://www.ftc.gov/news-events/blogs/business-blog/2019/07/d-link-settlement-internet-things-depends-secure-software> [Accessed 20 June 2020].
- Voas, J. K. R., Laplante, P. & Applebaum, S., 2018. *White Paper (Draft) : Internet of Things (IoT) Trust Concerns*. [Online] Available at: <https://csrc.nist.gov/CSRC/media/Publications/white-paper/2018/10/17/iot-trust-concerns/draft/documents/iot-trust-concerns-draft.pdf> [Accessed 20 December 2021].

New Wave Cyber Attacks

Angela Mison, Gareth Davies and Peter Eden

University of South Wales, Treforest, UK

Angela.Mison@southwales.ac.uk

Gareth.Davies@southwales.ac.uk

Peter.Eden@southwales.ac.uk

Abstract: There is increasing enthusiasm for, and recognition of, the benefits that artificial intelligence (AI) can provide to society. The emphasis has been on the positive, but AI and deep learning can be used for negative purposes. Modular Neural Networks (MNN) are capable of independent learning and have been targeted at evolutionary, complex financial systems. If the goal of an MNN were to be defined as system penetration, there is no reason why an algorithm could not run in the background. There are a resource requirements, but organised crime groups, technology companies, nation states and individuals with a curious bent are all capable of such. Ordered society and security requires a degree of certainty that systems on which society depends will remain recognisable, dependable and resilient. Under current conditions, security is difficult enough. It is suggested that limitations may be required before release of certain AI systems, in the knowledge of their potential for detriment to society. An AI system capable of independent learning, permits undefined emergent behaviours. That the results of any emergent properties may be benign or malign is irrelevant. Scientific history is littered with developments whose uses were redirected away from the benign. Such concern could be interpreted as fear of the unknown, standing in the way of technological advances. Unless society wishes to become machine-driven, the power and control of systems should be defined and limited by society, not accidentally sprung on humanity or based on a ruthless logic that may drive a system to an unacceptable conclusion. Currently there are sophisticated botnet forming methods ensuring botnet persistence. If combined with the concepts of AI, there is a possibility that botnets could exist in perpetuity, with no one able to predict emergence, and no time limits on evolution. Whither cyber defence in the face of the unstoppable, increasingly intelligent, goal directed systems?

Keywords: Goal directed AI, Bias, Emergent properties, Ordered society, Botnets, Persistence

1. Introduction

AI is an inspiring technology. It will be the most powerful tool in generations for benefitting humanity.....AI systems will also be used in the pursuit of power. We fear AI tools will be weapons of first resort in future conflicts. (National Security Commission on Artificial Intelligence, 2021)

Whilst the above is written from a Defence perspective, it is equally applicable to all systems as technological advances and research enable more attack techniques to move from Research or well-resourced Organised Crime Groups (OCG), to the DarkMarket, Crime as a Service and zeroday exploit corporates such as NSO Group and L3 Technologies. Indeed, the Final Report [op cit] even goes so far as to provide guidelines on how to hack AI systems.

With the exception of some command economies, in recognition that many organisations have neither the skill nor the capacity to undertake AI research or its implementation, it is anticipated that both protagonists and defenders will avail themselves of the talent within their respective equivalents of Silicon Valley or specialised software developers. This approach disregards the necessary aspect of expert customer and thus affects the specification of systems required, control of inherent bias, integration, the system roll out, policies and governance, and maintenance. Habitually, the person who understands the nuances, implications, and potential consequences is not the one at the top of the food chain, thus does not have the full influence over requirements, a relevance when budgets are involved and such concerns are likely to result in additional cost and possible extension of timeframe.

A description of managers in charge of something of which they know not provides evidence of the frustrations arising, as conveyed within the resignation post of the Chief Software Officer of the US Air Force (Chaillan, 2021). In his experience of chasing discontinuous funding, a coherent strategy for development of systems was undermined, at the same time as being drowned in bureaucracy. He bewailed the lack of agility in system development which kept systems current, relevant, and effective. While Chaillan [op cit] was speaking specifically of the Defence environment, the prevailing attitudes are recognisable to anyone from an IT projects background. The solution seems radical as it requires the cooperation and continuing involvement of a team of informed, expert commissioning parties.

The new wave cyber attacks are likely to be directed at data and communication paths, finding access through legacy systems, Cloud platforms, and corrupting AI algorithms to affect learning and alter goals. Data is essential for analytics and AI. A cautionary note could be applied comparing the model of AlphaGo Zero (Silver, et al., 2017) in which a system teaches itself by playing against itself, reinforcement learning, and the WOPR (War Games, 1983) in which an AI system ultimately concludes that all war scenarios are pointless, resulting in mutually assured destruction, and that the only way to win is not to start.

The National Security Commission on Artificial Intelligence [ibid] insists that whatever intelligent operational system is in use, there must always be a human in the loop. This does not necessarily hold outside the military complex, for example rights are given by the Article 22 of General Data Protection Regulation (European Parliament, 2016) to individuals to contest systems in which no human has participated. Presumably such cautions are to avoid a situation in which a learning system has corrupted goals and to counter irrationality of leadership that causes a declaration of war as in a film (Canadian Bacon, 1995), which plot bears remarkable similarities to some recent events.

2. The purpose of the AI system

It is perhaps more comprehensible to discuss purpose in a Defence context. The agility and flexibility it is thought to offer commanders in the field through the use of AI is misunderstood. For a system, any system, to be effective, it must have a defined purpose. For the military complex although on the battlefield it may seem obvious, to an AI developer, there is a need to define purpose and operational context. Without this definition, the AI developer is using his bias and knowledge to determine how to wage war and thus determining the behavioural actions and reactions of a programmed, intelligent device in the field.

The simple question, which perhaps no one wants to ask, is 'What is the purpose of this war?' Where this rises to the fore is in the use of autonomous devices. The rationale for the question is that a system will be created to wage war by evolving (AI) programmed boxes and the conduct of that war, which would normally be determined by the ethics and morality of the protagonists, is delegated and programmed into devices. What about encapsulation of United Nations ethics and morals, and alignment with international law?

Questions need answers. Is this war something that is being waged proactively as in the war on drugs, or something which is defensive, as against an attack? Does the behaviour differ in those circumstances? What are the rules of engagement? Do you choose the Rumsfeld, Bush, Powell, or any other doctrine? While it is possible to be more barbaric than your enemy and win a war, will that enable you to win the peace (Horne, 1978)? There is a concern that some nation states are in technological advance, but where they are not and it becomes a war of equals, does that lead to stagnation as in the First World War, or mutually assured destruction as in the Nuclear stand-off? In cyberspace, what are those equivalents and when will an incursion generate a retaliation?

There is much discussion of the cost savings and efficiencies surrounding repurposing AI systems. How do you repurpose AI, particularly after it has evolved in combat, learned and become something different from the original? Exactly what is it that you are repurposing? Furthermore, when repurposing, is this combat context the same as the previous context in which this system was used and evolved. Emphasis is placed on the human in the loop, but what happens when the human intervention runs counter to the behavioural choices of a device? Which one has ultimate control in an evolved system?

Diverting to civil authority purposes rather than the bellicose, surplus weaponry which has been purposed against combatants, such as the proposed taser carrying drones, can also be found in civil defence and police armouries enabling use, for example, in anti-terrorist/disaffected population control situations. Smith the inventor of the taser and body cam highlights:

Those groups are governed by different sets of rules and regulations; (Tucker, 2021)

Being governed by a different set of rules indicates that the goals may be different and therefore the system must behave differently. As an example, while lethal use of drone carried weaponry may be appropriate in anti-terrorist situations when the prime concern of authorities is protection of citizenry, in western liberal democracies, the public relations catastrophe associated with any such lethal use on popular demonstrations/riots defines that administration as no longer belonging to the class of democracies, but to the

class of autocracies. If in doubt, consider the international reaction to the authoritarian responses to demonstrations in Eastern European Democracies, Turkey, Hong Kong, Egypt etc.

In criminal hands, attaching any weaponry to drones is ripe for events such as drive by shootings, sentry duty, attacks on competitors and even attacking SWAT teams. It is much the same philosophy as use in military defensive capability, although the protagonists goals are different. OCG are exceedingly well resourced and funded through proceeds of crime. The profits from cybercrime are estimated to be at least equal to the Gross Domestic Product (GDP) of the Russian Federation.

[They] will continue to examine in an operational sense where cyber has utility, both in terms of the need to protect from it but also use it. It can be weaponised where appropriate against particular adversaries.
(British Army, 2021)

The above statement can apply to any OCG. They already have the capability to reverse engineer much of technology if they believe it can be used to their advantage and also may be major drivers of technological advance and a sink for well trained AI and cyber expertise.

3. Purpose and Cybersecurity

The answers to the above need to be programmed into devices, in advance and in the hope that your enemy will share your ethical standards. Part of the frustration expressed by Chaillan [*ibid*] was to do with recognition that in war, even a cold war, that may not be the case. It is assumed that one party to a conflict has a greater motivation to win, which leads that party to enact unethical behaviours - winning at all costs. It has been suggested that the inclusion of limiting factors in Stuxnet led to its attribution to western liberal democracy as no other states would have exhibited such an ethical approach.

Cybersecurity in its purest sense is about the security of systems. It is about protecting systems from externally induced corruption and ensuring their continuing operation in the manner specified. It is not about ethical and moral issues, nor the oversight of algorithm development. Systems are developed in accordance with requirements specifications. With intelligent systems, these are often goal directed and may contain various neural networks and be subject to machine or even deep learning capability. The way the systems adapt and grow is in response to a need to achieve their goal. There may be several behavioural steps on the way to achieving a goal, and identifying which one may be aberrant may be an impossibility for cybersecurity until malign behaviour is exhibited or a redefined goal is achieved.

It is easy to pick holes in the other people's work, but not particularly productive. A solution is required which is proactive rather than reactive. The initial risk assessment prioritises four areas: legacy systems, communications, data, and corruption of the learning mechanisms in AI systems.

4. Cybersecurity of legacy systems

Connected legacy systems are always problematic with respect to cybersecurity. They were written in the mists of time, like the 60 year old Inland Revenue Service (IRS) system holding taxation data of individuals. A time when it was believed the world was a more benign place. The USA IRS system will not complete its updating until 2030, when it will be replaced by a real time interactive system, under development since 2009 (Alms, 2021), thus already utilising obsolete systems and technology. It is difficult enough to find expertise and continuity of knowledge for maintaining the resilient operation of a legacy system, built in the times of early internet openness but this is compounded by trying to find a complete cybersecurity solution. For such systems, fortified cyber defence is possibly the only option, which provides unlimited access if breached. In an interconnected mesh, this is an undesirable characteristic.

The trial of linking legacy systems with interconnected systems is not just one of cyber security but also one of latency when data must be passed from one system to another. Without pattern of life information, under what circumstances can a change in latency be recognised as a cybersecurity failure. Too soon/early, it may be injected data, too late/slow, it may be an interface or communication failure. If it is too slow, what are the dependencies on the long awaited data which it is crucial to know for continuity purposes? If what is awaiting that data is an intelligent system, what happens?

With the passage of time, every system becomes a legacy system, requiring further patching and integration. A consideration, not necessarily directly related to pure cybersecurity, but nonetheless important, is the retention of legacy system skillsets. As a UK bank found to its cost, hiring back consultants you have just made redundant in a cost cutting exercise permits you to be held to ransom in an emergency. The moral of the UK situation is to only fire the experts after you have decommissioned the legacy system, its replacement is fully functional, and you are certain you have traced all the links to and from the legacy system and dealt with them. This latter is likely to be a growing problem in a mesh or interconnected system.

5. Cybersecurity of communications

There are many ways of disrupting communications. With trends towards utilisation of the Dew, Edge, and Fog, growing and increasing possibilities deriving from 5G and its successor 6G, evermore reliance is placed on a communications infrastructure. It is often seen to be advantageous to be at the forefront of technology - a pioneer, but as a tutor once said:

You can always tell the first adopters, they are the ones with the knife in their back. It is sometimes better to be a follower. (Sumner, 1984)

This seems to run counter to military need for flexibility and agility and business need for competitiveness, but it really is better to let someone else make the first mistakes.

Persistent decentralised botnets using peer to peer communication and updating whenever they encounter a newer version which can be changes at a rate that will evade the usual indicators of compromise can be an issue. These botnets can be used within the system or as an attack platform with a trusted source IP and all its concomitant implications.

Reverting to a military context, the relevance of, for example, the US Army's digital transformation program resulting in a mesh, which it is intended will connect the tactical edge to the enterprise and places all the data in the Cloud, sounds like great cost saving and efficient use of technology and infrastructure (US Army, 2021). However, it is a cybersecurity nightmare. It implies that there is a requirement for continuous communication, between and connection to all endpoints at potentially high volume.

There may well be a military equivalent of an autonomous vehicle's communication which is labelled V2X or vehicle to everything which adequately describes the risks associated with proposed IoT and Edge computing. Rather than subscribing to the secure by design paradigm, what is hinted at in (US Army, 2021) is the post hoc Line of Effort #3 which addresses cybersecurity and is exacerbated by subsequent demand for agility and flexibility in upgrading or otherwise changing the system.

The fact that the system is so ill-defined that there is a need to communicate with 'everything' means it is impossible to define the endpoints. The paradigm of endpoint security and zero trust as espoused in Secure Beyond Breach (Reiber, 2020) provides insufficient protection in an open interconnected system where endpoints cannot be defined. The adoption of microsegmentation of networks requires a system boundary definition, identification of the virtual networks operating within that context and consideration of white lists. These very acts constrain the development of systems. Any component system must contribute, in some defined way, to the higher goals identified - its purpose. If something does not contribute to the goal, it has no place in the system. Whether it is necessary for a battlefield commander to have connection and data transfer between the battlefield and the enterprise should be a subject for further discussion.

The reliance on AI and Cloud technology places an essential burden on communications which must be available 24/7/365 with no down time where there is a dependence on real time data. In business, a breakdown in cybersecurity of technological communications becomes an operational continuity problem and there is a long list of documentation necessary to deal with such a situation, contained in the Blue Team Field Manual (White & Clark, 2017).

If security of communications is breached, there should be a continuity plan. However, this plan is not the sole responsibility of the Incident Response Team. The plan identifies the critical operational plans, and in the height of battle, there may be no requirement for enterprise information. Such a distinction implies that there may be a number of continuity plans based on specific scenarios. The proposed mesh may break, but a distinction

between a real time operational platform and a less time critical enterprise platform dictates the priority of response and perhaps provides some network design input.

In simple terms, if there is building security and the building is evacuated due to some event, would it be expected that the building security team would dictate operational continuity of the organisation and that each element of the operation was of equal importance? Stangely, the operational continuity plan seems to fall into cybersecurity responsibility, and be accepted by them as part of their responsibility for incident response. It contributes to the skills shortage, distances personnel from operations and curtails the cybersecurity aware culture. Prevailing attitudes may become more entrenched as dependence on systems and distance from them increases.

6. Cybersecurity of data: the Cloud

There are the usual attacks on data in transit available, to which it may be assumed that encryption is the answer. This is not always the case in time/mission critical systems, which may include AI, where resource at the operational/functional end is limited. The processing overhead is not always acceptable or feasible.

Army CIO has said the Army does not have

the right policies and procedures for some recent technologies, such as internet of things devices, or ways protect them on DOD's network (Williams, 2021).

though the National Institute of Standards and Technology (NIST) is addressing part of the issue, pursuant to an Executive Order (Biden Jr, 2021). Exactly the same sentiment could be expressed by any organisation pursuing digital transformation. The question must be asked that with finite resources what is the priority, and the answer is actually something that was not mentioned, data. If the network was intact and entire, it would be useless unless information/data are flowing through it.

Part of the reason for outsourcing to the Cloud is a cybersecurity skills shortage together, a belief that a dedicated, expert cybersecurity team is available to the Cloud Service Provider (CSP), and a potentially flexible infrastructure. The CSP has the knowledge of the infrastructure and responsibility for its cybersecurity. Unfortunately the Cloud is not inviolate from attack, and there may not be restrictions on its location or movement of data between one location and another. Where security is based on pattern of life, such movement may disrupt any latency definitions, raising alarm where it is not necessary. Cybersecurity may not be able to differentiate. There may be cybersecurity failures within the Cloud itself and unless the client arrogates investigative rights, it could well remain hidden or incompletely rectified.

Moves are afoot in the US, and presumably elsewhere, regarding regulation of access and audit clauses to be contained in contracts with CSP. Such considerations apply equally to any third party attached to the interconnected network and to any third party componentry, hardware, firmware or software as each represents an opportunity to a malefactor. There is the also the dangerous belief in the invincibility of one's own cybersecurity the initial belief that any failure must derive from elsewhere in the interconnected mesh.

In conclusion, recognition of issues concerning data, communications and the Cloud do not stop with cybersecurity. As Army CIO Iyer stated:

We know how to patch traditional IT systems. How do you patch sensors and operational technology, how do we collect the right analytics and data, and how do we share data with our joint service partners and industry. That is a policy problem (Williams, 2021)

And everyone knows how long policy problems take to resolve.

7. Cybersecurity of learning mechanisms

As indicated in the introduction, corruption of learning systems may be the goal of any adversary or competitor. Subversion of a system would provide an advantage to the perpetrator. It is not certain how such an intelligent system could be fixed. How is the exact point of corruption found and the system unwound? In the case of a goal directed intelligent system unleashed onto the internet, could it be retrieved?

It is not clear how aberrant behaviour could be guaranteed to be detected, let alone proactively prevented unless there exists a boundary definition, a clear set of operational parameters and a unique pattern of life in execution. Evolving systems are not amenable to pattern of life mechanisms as the system is in a state of continuous 'growth' or evolution.

Again, as with the autonomous vehicle, the learning element could be carried out on 'land based' computers defining the limits of acceptable behaviours which contribute to the goal and are in accordance with permissible and possible behaviours and a device's capability. Once the scenarios have been defined, and system learning undertaken, the resultant predictable selection could be uploaded as an immutable system to the requisite device. Obviously this lacks the flexibility of military expectations, but permits more of a plug and play capability with each system learning from downloaded previous 'experience', adapting to those experiences to produce a further predictable system which is uploaded to the device in its entirety.

Uncontained intelligent systems can lead to unpredictable emergent properties and behaviours whereby the system as a whole behaves in a way that is not foreseeable from its collection of parts. A biological example of this is the collection of neurons and synapses which effectively form an electrical circuit, which is the basis of the brain. The brain has the emergent property of situational awareness. None of the individual component pieces possess this emergent property, but in its entirety, the brain does. The time of its emergence, functionality and any further evolution is not predictable. It is not possible to predict whether the emergent property will be malign or benign, useful or not.

There are potential legal implications for AI systems. These have not yet been tested. The Courts are demanding trustworthy AI, or at least a comprehensible explanation of why and how a system has behaved in a particular manner and whether that behaviour is reasonable. Some governments are demanding audits of algorithms to determine if they are equitable or have developed from an initial bias. This audit would take specialist training and it is not certain that it is within the cybersecurity remit although it must, certainly, be part of a risk assessment. The mechanism for the justification of Facebook algorithms which recognise anger, react to it, and reinforce or amplify it will be an interesting test case addressing as it does bias and misinformation, although these are tied together with surveillance capitalism (Zuboff, 2019) and analytics.

There is some thought to declaring AI systems to be legal entities in their own right (Abbott, 2020) and at this, the role of cybersecurity becomes even less certain.

8. Conclusion

This has been a diversionary discussion of evolving technology and issues concerning that evolution, clickbait in effect. No system is immune from attack and many organisations have the capability and time to invest in developing techniques to achieve a hack.

With the increase in interconnectivity and openness, there are limits to what cybersecurity can achieve. The adage 'a system's security is only as strong as its weakest link' prevails. What has emerged is a need to consider the true role of cybersecurity professionals, the realistic limits of their tasks and the systems they are trying to secure.

A recognition that the security associated with increasing use of AI systems may not be the preserve of current cybersecurity professionals forces decisions concerning definitions of purpose, ethics, morals, bias and behaviour to further professional classes (DeepMind, 2021). Currently it may not be possible to provide the full flexibility of AI systems that a military force might desire/need safely. With respect to unconstrained learning systems, even Google terminated the program for AlphaGo Zero before it had reached its full intelligent capability, so its full capability will remain unknown.

References

- Abbott, R., 2020. *The reasonable Robot: Artificial Intelligence and the Law*. Cambridge: Cambridge University Press.
- Alms, N., 2021. A 60 year-old IRS IT system won't finish modernising until 2030. [Online] Available at: https://fcw.com/articles/2021/10/20/irs-gao-cade2-delays.aspx?oly_enc_id= [Accessed 21 October 2021].
- Biden Jr, J., 2021. Executive Order on Improving the Nation's Cybersecurity. [Online] Available at: <https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/> [Accessed 15 October 2021].

- British Army, 2021. THEIA: The British Army's Digital Transformation. [Online] Available at: <https://www.purestorage.com/content/dam/pdf/en/misc/british-army-digital-transformation.pdf> [Accessed 15 October 2021].
- Canadian Bacon. 1995. [Film] Directed by Michael Moore. USA: Metro Goldwyn Meyer.
- Chaillan, N., 2021. It's time to say Goodbye!. [Online] [Accessed 12 October 2021].
- Davies, G., 2021. Private Conversation [Interview] (4 June 2021).
- DeepMind, 2021. Safety and Ethics. [Online] Available at: <https://deepmind.com/safety-and-ethics> [Accessed 15 October 2021].
- European Parliament, 2016. Regulation (EU) 2016/679. [Online] Available at: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> [Accessed 15 October 2021].
- Horne, A., 1978. A Savage War of Peace: Algeria 1954-1962. New York: Viking Press.
- National Security Commission on Artificial Intelligence, 2021. Final Report, Washington DC: United States Government.
- Reiber, J., 2020. Secure Beyond Breach. [Online] Available at: <https://www.illumio.com/sites/default/files/2021-02/micro-segmentation-secure-beyond-breach-20eb10%20%281%29.pdf> [Accessed 13 October 2021].
- Silver, D. S. J. et al., 2017. Mastering the game of Go without human knowledge. [Online] Available at: https://www.nature.com/articles/nature24270.epdf?author_access_token=VjXbVjaSHxFOctQQ4p2k4tRgN0jAjWel9jnR3ZoTv0PVW4gB86EEpGqTRDtpIz-2rmo8-KG06ggVobU5NSCFehILHcVFUeMsbvws-lxjqQGg98faovwixeTUgZAUMnRQ [Accessed 15 October 2021].
- Sumner, F., 1984. Private conversation [Interview] 1984.
- Tucker, P., 2021. The inventor of the Taser and the Body Cam Wants to put them on Drones. [Online] Available at: <https://www.defenseone.com/technology/2021/10/inventor-taser-and-body-cam-wants-put-them-drones/186095/> [Accessed 15 October 2021].
- US Army, 2021. The Army Unified Network Plan 2021 Enabling Multi Domain Operations. [Online] Available at: <https://api.army.mil/e2/c/downloads/2021/10/07/d43180cc/army-unified-network-plan-2021.pdf> [Accessed 15 October 2021].
- War Games. 1983. [Film] Directed by John Badham. USA: United Artists; Sherwood Productions.
- White, A. & Clark, B., 2017. Blue Team Field Manual. version 1.2 ed. Scotts Valley(California): CreateSpace Independent Publishing Platform.
- Williams, L., 2021. ARMY CIO sets out to revamp IT policy. [Online] Available at: https://fcw.com/articles/2021/10/14/iyer-army-tech-policy-revamp.aspx?oly_enc_id= [Accessed 15 October 2021].
- Zuboff, S., 2019. The Age of Surveillance Capitalism: The fight for a human future at the new frontier of power. 1st ed. London: Profile Books.

The Role of Big Tech in Future Cyber Defence

Angela Mison, Gareth Davies and Peter Eden

University of South Wales, Treforest, UK

Angela.Mison@southwales.ac.uk

Gareth.Davies@southwales.ac.uk

Peter.Eden@southwales.ac.uk

Abstract: Ordered society and nation states are dependent on interconnected systems, the defence of which is largely in private hands whose actions are driven by need for oligopolistic market dominance, protection of assets, and their monetisation models. This paper queries the responsibility of the nation state for the protection of itself and its citizenry. By some definitions, corporations are conducting cyberwarfare and, in cyberspace, are virtual nation states with ownership and rights over the data they hold and the intelligence it yields. The financial challenge for market dominance could drive an internecine war among the major technology corporations, and an assertion that the rights over the data they control are superior to those of the nation state. As functional monopolists, data they have acquired is not available from any other source. The intelligence from analytics exercised over that data, and the data itself is proprietary. These corporations exercise monopolist characteristics in the areas of data, information and intelligence. The aggregate value of the top 5 technology corporations, colloquially known as Big Tech is equivalent to third in projected global GDP rankings for 2021. This represents an equivalent expression of power in/over cyberspace. Cloud service providers (CSP) are often offshoots of Big Tech and have a high compound annual growth rate, thereby revealing the motivation for protection of market dominance and potential threat to user/customers. By concentrating on traditional cyber warfare and defence, there is limited consideration on policing or guarding against the rise of these virtual supranational powers driven by strict market agenda. What consideration there is regarding potential threats is driven by an economic perspective and anti-trust initiatives. Whether judged by the nation state as benign or malign, Big Tech has an impact on the nature and direction of society as currently understood and the question must be raised whether both citizens, organisations, and states need protection from it.

Keywords: Big Tech, Economic power, Surveillance capitalism, Information warfare, Cybersecurity

1. Introduction

1.1 The economic position of Big Tech

To the nation state seeking information, Big Tech are oligopolists for data, information and intelligence. Their motivations are strictly financial with respect to market dominance. Their extra-jurisdictional capacity enables them to put data beyond the reach of nation states in uncooperative jurisdictions, without the potential for retaliation, due to the dependence of nation states on the very organisations they may be seeking to regulate. Furthermore, their oligopsonistic position endangers and influences the direction of technological development, education, and exacerbates the skills shortage as their demand for specialist expertise increases.

Any form of market dominance can be detrimental to both customers and suppliers. Governments attempt to regulate the size and structure of market dominant organisations, requiring them to act independently of each other, and to promote competition for goods and services at beneficial prices (The Free Dictionary by Farlex, 2021). Historically, such anti-trust examples can be demonstrated by the break up of Ma-Bell in the United States of America and regulation of Microsoft in Europe.

1.2 Information warfare

There are many definitions of warfare in the digital sphere. Most mention the nation state and assume a political motivation, interfering with and potentially damaging information systems of the attacked entity. The suggested internecine war for economic market dominance could be seen in the light of the following definition where the battlefield is represented by the customer base and the combatants are big tech:

Information warfare is combat operations in a high-tech battlefield environment in which both sides use information technology means, equipment, or systems in a rivalry over the power to obtain, control, and use information (Wang & Li, 1995)

2. Is Big Tech an Oligopoly

There are technical economic definitions of market dominance for both the supply side and purchasing side. Oligopoly which relates to the supply side, describes a market in which there are a small number of players

present. Their products are closely related. It represents a competition among the few. This can be detrimental to customers/users as players can dictate both terms and prices. Where a customer is dependent on the good or service provided, the player can exercise undue or coercive influence over him.

Oligopsony is a purchasing market equivalent. Due to their purchasing power oligopsonists can dictate terms, prices, what is produced and where. Oligopsonists can have a collateral effect on other areas e.g. education for training to fulfil their needs or the direction and pace of research and development.

There is an argument that Big Tech does not represent an oligopoly, but that the constituent organisations operate in different markets. Where they compete, they are differentiated by behaviour and business models (Lotz, 2018).

It is suggested that Facebook and Google between them control 63% of the digital advertising market, and that advertising contributes to 97% and 88% of their revenues respectively. Both target consumers through algorithms. Their models of surveillance capitalism are well tried and profitable due to successful micromanipulation of users (Committee on the Judiciary, 2021).

Google achieves the bulk of its revenues from searches and has a pay per click business model, while Facebook uses gratis generation of attention-grabbing content and has a pay per view business model. Google makes a direct comparison of advertising revenue generated between its own search engine and Microsoft's Bing (Johnson, 2021), indicating a monitoring of potential challengers for market dominance.

There is a growing base of users of Google Cloud Platform and, using it as an example, it advertises 222% return on investment within 3 years and a payback on investment of 8 months. It is comprehensible that, in addition to other cost efficiencies (Lava & Marden, 2020), CSP are experiencing unprecedented growth.

Apple gains 84% of its revenues from hardware sales, where its profit margin is so high that market domination is unnecessary. It is a vertically integrated organisation, a silo, whose products are highly integrated and optimised when operating on its platforms. It is a growing platform conglomerate in that 11% of its revenues are generated through access to goods and services accessible through its platforms, e.g. media outlets and payment mechanisms (Vailshery, 2021).

Amazon is primarily a retailer, gaining 70% of its revenues directly from retail sales with 200 million unique visitor users per month (Hufford, 2020). For other retailers, Amazon acts as an intermediary. Businesses benefit through brand association. It is a platform conglomerate (AWS).

16% of AWS users are classified as large enterprise, while 84% represents Small and Medium sized enterprises, to whom outsourcing security and operational expertise, is part of addressing the skills shortage, cutting costs, and reducing their risk exposure. In the same manner as Google Cloud, Amazon sells the business value of AWS directly to the Board (Amazon Web Services, 2021). It does has sales documentation addressing cybersecurity (Rodgers, 2021), pertaining to the commissioning organisation, but not the security practices and procedures within AWS.

Microsoft predominates in the sale and licensing of software and, more recently, has entered the hardware market. Microsoft has an expansion strategy encompassing Cloud offerings, Azure. Azure is thought to have 4 Million customers globally (InfoClutch, 2021).

With the exception of Apple, a niche player, it is possible to agree that Big Tech is not necessarily an oligopoly, although they behave as such, and meet the criteria for oligopsony in their markets. While there have been concerns over the anti-competitive behaviour of Microsoft and Apple, they also demonstrate that there are advantages to oligopolistic supply and oligopsonistic diktat. The degree of interoperability available between disparate manufacturers is in part due to the running of the Microsoft operating system. The effective splitting of the mobile phone market between Android and Apple iOS epitomises the difficulty in achieving a balance between the pros and cons of any anti-trust action.

3. The absolute economic power of Big Tech

A country's comparative wealth ranking is usually made through use of Gross Domestic Product (GDP), the total value of all goods made, and services provided during a specific time period (HM Treasury, 2017). Corporate comparisons use market capitalisation. For the purposes of this section, the global GDP rankings, except for North Korea, are taken from (World Bank, 2021) while the market capitalisation for Big Tech are taken from (CompaniesMarketCap, 2021)

Currently, Big Tech comprises 5 USA based corporations. Taken individually, Big Tech corporations would rank between the fifth and sixteenth countries in the global GDP rankings. The aggregate market capital of the 5 organisations is \$9.2 Trillion, roughly equivalent to the third largest global economy after the United States of America (GDP \$20.9 Trillion) and China (GDP \$14.7 Trillion). In comparison, the countries providing the greatest cybersecurity threat in addition to China are the Russian Federation (GDP \$1.5 Trillion) and the Democratic Peoples' Republic of Korea (GDP \$18 Billion), (Trading Economics, 2021). Chinese Big Tech, Tencent (\$558 Billion), Alibaba (\$407 Billion), and Baidu (\$55 Billion) (Companies Market Cap (A), 2021) exhibit geographic market dominance.

Oligopsonistic power exercised over technology manufacturers, researchers, developers and education is revealed by consideration of the aggregate economic power of a purchaser with the world's third largest economy. Most recently the direction of the AI research may be subject to influence through funding provided to Universities by private corporations.

4. Big Tech and the growth of Cloud Services Providers

4.1 Market size, share, and trends

The economic power of Big Tech represents an equivalent expression of power in/over cyberspace. Big Tech has offshoot CSPs. A report on Cloud computing market size, share and trends (Grand View Research, 2021) provides the following details:

CSP are estimated to have a compound annual growth rate (CAGR) of 19.1%. With such a statistic, it is obvious that there is motivation for achievement and protection of market dominance. Big Tech has the resources for such competition. The global client spend on Cloud Services is estimated to be \$1 Trillion, with a CAGR of 15.7%. The transition to the Cloud has accelerated due to the Covid-19 pandemic and the move to flexible and remote working.

For the individual consumer, the bulk of growth is seen in the Cloud Storage market where the major players are Apple, Alphabet (Google), Amazon, Microsoft, Box and Dropbox. The Cloud Gaming market is cross platform with an estimated 0.5 Billion users and a CAGR of 48%.

For businesses, the greatest area of usage for CSP is for business intelligence, supply chain management, enterprise resource planning and project and portfolio management. In terms of CSP revenue, X as a Service accounted for 60% of the revenue stream.

4.2 Cybersecurity and surveillance capitalism

What should not be forgotten in terms of cybersecurity is the balance of CSP revenue streams attributable to surveillance capitalism. Knowledge of an organisation, mode of operation and data gained through surveillance capitalism could be viewed as equivalent to an organisation's intellectual property, yielding the opportunity for such intelligence to influence negotiations with the CSP. Given the level of insider knowledge, this knowledge could be used by Big Tech to expand via merger and/or acquisition, increasing further their economic power. The business opportunity for OCG to expand their private Cloud operations to offer 'legitimate' Cloud services as a mechanism for money laundering through investment in legitimate businesses and opportunities for growth through use of such insider information is manifest.

Given the slick marketing of the business value of Cloud Services to the Board, it is easy to see that any cybersecurity technical queries raised with the Board may not carry sufficient weight to mitigate consequential outsourcing risks. The details of cybersecurity policies and protections applied internally to the CSP are, not without reason, a closed subject.

4.3 Oligopsony, oligopoly, and interoperability

The projected market dominance enables CSP to act as oligopsonists to their infrastructure suppliers, software suppliers, and academia, dictating both the direction and speed of technological development, advances in artificial intelligence and analytics, and the concomitant education required. The employment market also demonstrates this economic power. Organisations may have to accept the continuous provision of training positions as experienced personnel migrate to Big Tech with whom they cannot compete.

Competition and differentiation is already developing between the Big Tech constituents as they tussle for market dominance introducing a no return scenario for customers migrating to them. It is unlikely that any government or global organisational spend will be of sufficient quantity to influence the direction of Big Tech. Without interoperability, nations' dependence on CSP leaves them open to pressure and manipulation. For Big Tech, the fight for market dominance as CSP provides the battlefield for information warfare.

5. The costs of cyber defence and cybercrime

Robert Mueller III, a former FBI Director stated:

There are only two types of company: Those that have been hacked and those that will be hacked (Mueller III, 2012)

There is an estimated global cybersecurity workforce shortage of 3.5 million (Moore, 2021) in a context where the global cost of cybercrime, \$5.5 Trillion in 2020, is projected to rise to \$10.5 trillion in 2025 (Nai, Fovino et al., 2020). Compared to GDP, the 2025 cost equates to the third largest global economy. Cybersecurity markets have a 2020 value of \$200 Billion with a CAGR of 10% (*op cit*).

Revenues from cybercrime are of the order of \$1.5 Trillion, 1% of global GDP, reputed to be greater than the profits of the global drug trade (McGuire, 2018), equivalent to the Russian Federation GDP (*ibid*). It has been suggested that some nation states benefit directly from cybercrime revenues and there are close links with money laundering (Office of Director of National Intelligence, 2021), (Miralgia, Ochoa et al 2012). With increased legalisation of drugs threatening to reduce organised crime group (OCG) profits, the logical business decision for OCG is to diversify into safer cybercrime with its low cost of entry and high return. To fully understand the volumes of money flowing through OCG, during the financial crisis of 2008, OCG is reputed to have supported the global economy (Spannaus, 2012).

Subcontracting by OCG could lead to an increase in Crime as a Service (CraaS) and introduce plausible deniability. McGuire [*ibid*] suggests that platform capitalism has been the intermediation model for CraaS. With revenues of \$1.6 Billion, CraaS comprises sub and prime-contracted groups akin to a guerrilla army. The financial and human economics of guerrilla warfare are well known, as is the cost of defence against opportunistic cyber-attack. The situation portrayed in (Scott, 2004) has amplified with time.

In addition to the suggested business expansion schemes (Global Initiative Against Transnational Organised Crime, 2021), OCG have the means and patience to recruit, train, and retain the best, often behind legitimate business facades. A reported reinvestment programme of \$300 million financed by revenues of cybercrime allows for investigation and development of new crimes, research on approaches not yet countermanded, and novel techniques (National Crime Agency, 2021). Defence against well-resourced, and motivated opponents is high cost. OCG cybercrime may represent a further example of information warfare involving market dominance.

6. The nation state, cybersecurity, information warfare, and anti-trust initiatives

Big Tech are using information technology, equipment and systems to attain their end - market dominance. Using the definition of information warfare [*ibid*], this maps directly to the rivalry over the power to obtain, control, and use information.

It is a State's responsibility to ensure an ordered society and the protection of its citizenry from all forms of warfare. Organisations, individuals, and even nation states are seemingly unaware of being victims in an internecine war, playing a very unequal part, and being on the losing side as they offer themselves and their information freely in exchange for business efficiencies and functionality.

Ordered society is dependent on the continuing and predictable behaviour of interconnected systems and the critical national infrastructure. The defence of such systems is largely in private hands whose levels of expertise, and knowledge of hyperconnected systems environment and complexity is unknown. In transitioning to the Cloud, elements of cybersecurity are being outsourced with no guaranteed access to policies, governance and procedures of providers and no inherent right to any data or the intelligence generated via analysis of that data.

The government can mandate what protections should be in place to ensure the continuity of operation of such systems, but a distinction can be drawn between compliance and performance. Technological advances mean that the knowledge for and complexity of securing any hyperconnected system cannot be determined, when a component is Cloud based. CSP are singularly reticent concerning their own cybersecurity, but Mueller's words [*ibid*] are apposite. If what organisations are protecting or securing cannot be defined, it increases the likelihood of cybersecurity failure. Such is the dependence on systems, that continuity plans are often sparse and/vague.

If governments, other organisations or individuals contract with CSP, legal entities in their own right, those CSP can relocate their corporate registrations and data repositories to the equivalent of cyber havens, rendering themselves extra-jurisdictional. In the worst case, uncooperative or seemingly over-regulatory governments could find themselves trapped in a technology without the portability and interoperability to enable them to transfer to another provider without interruption.

Anti-trust initiatives should be considered in this light. If Big Tech is broken up, the responsibility for continuity of operation of CSP must be specified. Contracts should include clauses concerning arrogation. The lack of available expertise for maintenance of the systems thus seized, is problematic, but seizure could extend a transition and migration period. Otherwise, undue pressure may be brought to bear on legislators or regulators for a less than optimum solution which exacerbates the potential for oligopolistic or oligopsonistic abuse.

7. Big Tech and data

Ownership of data is defined according to jurisdiction. Ownership of and rights to data is a particularly thorny issue that requires a paper in its own right (Scofield, 1998) (Northern Illinois University, 2021). There are two aspects of Big Tech and data that will be touched on here: the individual and the requirement for legitimate access to data by the nation state.

There is a United Nations initiative promoting data sovereignty, (Internet Governance Forum, 2020) which has been taken up by one of the internet founders (Berners-Lee, 2021). Such a move, giving individuals absolute right over their personal data and removing it from the control and ownership of Big Tech only partially addresses the problems caused by the growth of surveillance capitalism.

A possible extension to data sovereignty is the Swiss initiative examining the potential for the unification of online and offline presence to define an individual [*ibid*]. An online presence can be interpreted as an avatar, similarly to personality dissociative disorder. The individual becomes the sum his parts. The extension to unification of the person includes metadata. Metadata is part of the individual as it is data arising from actions and behaviours online. Such unification of the online and physical manifestations of the person, provides the state with investigative rights over its compound citizens.

Removal of such a major revenue stream arising from data would reduce the economic power and valuation of Big Tech and copious other industries. Extreme resistance can be expected. It could herald the advent of paid applications as organisations sought to recover losses. In the argument over privacy versus functionality, it is often stated there is '*no such thing as a free lunch*' (Heinlein, 1966).

While data sovereignty of the individual may ease the path of intelligence services and law enforcement agencies by giving them a quick route to some data access, what is sometimes of more interest and more informative are metadata and management data. Metadata arises from the system itself, thus there is an argument that with pure data sovereignty, there is no relation between the individual and data generated within the system.

Management data, useful inter alia for demand prediction and billing, is indubitably the property of the CSP, but may be extremely relevant with regard to an investigation. Access to strictly proprietary data may be difficult to argue, and its link to criminal investigations bears a resemblance to the proactive preventive activities of

Minority Report (Dick, 2017). Of note in this area, is the treatment of data requested to be retained and later demanded by the Congressional Committee investigating events of January 6 2021 and the legal arguments associated with it (Goltein 2021).

8. Big Tech, policy, law, intelligence and law enforcement

For interconnected systems there is a difference between physical location and system accessibility. With growing economic power and capacity to relocate, Big Tech can be considered supranational entities with jurisdictional issues arise over access to data (Mison, 2019). The entities involved in CSP operation and their location are often private to the CSP. There are some similarities between CSP and the banking industry. The nation state is hampered by jurisdiction and international law. Data is a highly mobile commodity.

There is no law governing cyberspace nor the creation of supranational entities. Policy makers should consider the possible fallout from information warfare and have requisite policies and statutes in place. Technological development, policy definition, and statute enactment move at radically different speeds. There is a need to anticipate cybersecurity needs. If the guess about the future is wrong, having the defence is not an issue other than cost, but not having it, could lead to chaos.

For nation states dependent for their continuing operation on Big Tech CSP, containing the power of Big Tech may be limited by that dependence. Due to inherent market dominant behaviour, and limitations of portability and interoperability, a major cybersecurity concern for the existing new and forthcoming digital transformation programmes is to avoid being locked into a platform or service. Without an acceptable means of exit, Big Tech can exert extreme pressure, forcing administrations to yield to their demands.

The importance of understanding outsourcing cybersecurity to the CSP leads to internal CSP cybersecurity failures and the consequential impact on clients and beyond. Notification is at the whim of the CSP and investigation is contained within it, with no rights for any independent authority which may exist. Where there is any dependence, there should be contractual obligations placed on the CSP that enable independent investigation and assessment by the nation state or its agencies. Such cybersecurity outsourcing exposes dependent systems to vulnerabilities on a scale not hitherto seen as evidenced by log4j CVE-2021-44228 (mitre.org, 2021).

9. Conclusion

Nation states concentrate on traditional cyber warfare and cyber defence in order to protect themselves and their citizenry from the machinations of others. Big Tech represents a different threat, a clear and present danger. Already there is increased complexity of interconnected systems, increasing cybersecurity and cyber defence costs. If Big Tech and CSP approximate to supranational powers, there will be a need to police and guard against the imposition of their market driven agendas. One of those vying for market dominance in the future may just be an OCG legitimate facade e.g. a VPN or CSP.

To assess the future impact on society of unconstrained and uncontrollable Big Tech, one need look no further than the American experiment and its results which have led to self-perpetuation and reinforcement of strong, radical division threatening democracy (Egan, 2020). Whether judged by the nation state as benign or malign, technology has a great impact on the nature of society as it is currently understood and the question must be raised whether they should be considered as an integral factor in the formation of any nation's or organisation's cyber defence plans.

References

- Amazon Web Services, 2021. *Realizing Business Value with AWS*. [Online] Available at: <https://aws.amazon.com/executive-insights/content/realizing-business-value-with-aws/> [Accessed 22 September 2021].
- Berners-Lee, T., 2021. *Time Berners-Lee's plan to save the internet: give us back control of our data*. [Online] Available at: <https://theconversation.com/tim-berners-lees-plan-to-save-the-internet-give-us-back-control-of-our-data-154130> [Accessed 22 September 2021].
- Committee on the Judiciary, 2021. *Algorithms and Amplification: How social media platforms 'Design choices shape our discourse and our minds* [Online] Available at: <https://www.judiciary.senate.gov/meetings/algorithms-and-amplification-how-social-media-platforms-design-choices-shape-our-discourse-and-our-minds> [Accessed 3 December 2021]

- Companies Market Cap (A), 2021. *Largest Chinese companies by market capitalisation*. [Online] Available at: <https://companiesmarketcap.com/china/largest-companies-in-china-by-market-cap/> [Accessed 27 September 2021].
- CompaniesMarketCap, 2021. *Largest Tech Companies by Market Cap (B)*. [Online] Available at: <https://companiesmarketcap.com/tech/largest-tech-companies-by-market-cap/> [Accessed 27 September 2021].
- Dick, P., 2017. *Minority Report: Volume Four of the Collected Stories*. London: Orion Publishing Group.
- Doyle, C., 2002. *Market Definition and Dominance*. [Online] Available at: <https://www.itu.int/osg/spu/ni/competition/Presentations/Market%20Definition%20CDoyle.pdf> [Accessed 22 September 2021].
- Egan, PJ, 2020. *Elections 2020 – The Role of Social Media in US Elections* [Online] Available at: [Elections 2020- The Role of Social Media in U.S. Elections - United States Department of State](#) [Accessed 30 November 2021]
- Global Initiative Against Transnational Organised Crime, 2021. *Organised Crime Index*. [Online] Available at: <https://ocindex.net/> [Accessed 30 November 2021].
- Golstein E, 2021. *Congress’ access to individuals private communications: the Jan 6 Committee’s troubling precedent* [Online] [Congress’ Access to Individuals’ Private Communications: The Jan. 6 Committee’s Troubling Precedent \(justsecurity.org\)](#) [Accessed 30 November 2021]
- Grand View Research, 2021. *Cloud computing market size, share & trends analysis report by Service, by enterprise size, by end use, by deployment and segment forecasts 2021 - 2028*. [Online] Available at: <https://www.grandviewresearch.com/industry-analysis/cloud-computing-industry> [Accessed 22 September 2021].
- Hattangandi, V., 2017. *Market Dominance Strategies*. [Online] Available at: <https://drvidyahattangadi.com/market-dominance-strategies/> [Accessed 22 September 2021].
- Heinlein, R., 1966. *The Moon is a Harsh Mistress*. 1st ed. New York: Putnam.
- HM Treasury, 2017. *Gross Domestic Product (GDP): What it means and why it matters*. [Online] Available at: <https://www.gov.uk/government/news/gross-domestic-product-gdp-what-it-means-and-why-it-matters> [Accessed 22 September 2021].
- Hufford, J., 2020. *Amazon Statistics: Need to know numbers about Amazon*. [Online] Available at: <https://www.nchannel.com/blog/amazon-statistics/> [Accessed 22 September 2021].
- InfoClutch, 2021. *Installed base*. [Online] Available at: <https://www.infoclutch.com/installed-base/cloud-computing-software/microsoft-azure/> [Accessed 22 September 2021].
- Internet Governance Forum, 2020. *IGF 2020 #42 Personal Sovereignty: Digital Trust in the Algorithmic Age* [Online Discussion] Available at: <https://www.intgovforum.org/en/content/igf-2020-of-42-personal-sovereignty-digital-trust-in-the-algorithmic-age-0> [Accessed 30 November 2020].
- Johnson, J., 2021. *Google vs. Microsoft Advertising: US mobile click share 2021*. [Online] Available at: <https://www.statista.com/statistics/223308/trend-in-impressions-on-google-and-bing-in-the-us/> [Accessed 22 September 2021].
- Lava, S. & Marden, M., 2020. *The Business Value of Improved Performance and Efficiency with Google Cloud Platform*. [Online] Available at: https://services.google.com/fh/files/misc/idc_business_value_of_google_cloud_platform_whitepaper.pdf [Accessed 22 September 2021].
- Lotz, A., 2018. *Big Tech isn't one big monopoly it's 5 companies all in different businesses*. [Online] Available at: <https://theconversation.com/big-tech-isnt-one-big-monopoly-its-5-companies-all-in-different-businesses-92791> [Accessed 22 September 2021].
- McGuire, M., 2018. *Nation States, Cyberconflict, and the Web of Profit*. [Online] Available at: https://images.marketingcontent.ext.hp.com/Web/HPMartech/%7Bbbdbdbcf-1506-4a26-aaa2-db614096130e%7D_hp-bps-web-of-profit-report_APR_2021.pdf?elqTrackId=2756f4161eb1467e98921ad587b6e0b7&elqaid=436&elqat=2 [Accessed 22 September 2021].
- Merriam-Webster, 2021. *Definition of Monopsony*. [Online] Available at: <https://www.merriam-webster.com/dictionary/monopsony> [Accessed 22 September 2021].
- Miraglia, P; Ochoa, R; Briscoe, I; 2012 *Transnational Organised Crime and Fragile States* [Online] Available at: [WP3 Transnational organised crime.pdf \(oecd.org\)](#) [Accessed 30 November 2021]
- Mison, A, 2019. *MSc Dissertation: Defining a method for successful use of evidence obtained from Cloud based data in prosecutions in England and Wales* University of South Wales
- mitre.org. 2021. *CVE – CVE-2021-44228* [Online] Available at: [CVE - CVE-2021-44228 \(mitre.org\)](#) [Accessed 13 December 2021]
- Moore, M., 2021. *Cybersecurity Employment Growth Report*. [Online] Available at: <https://onlinedegrees.sandiego.edu/cybersecurity-jobs-report/> [Accessed 22 September 2021].
- Mueller III, R., 2012. *Speeches: RSA CyberSecurity Conference, San Francisco, CA*. [Online] Available at: <https://archives.fbi.gov/archives/news/speeches/combating-threats-in-the-cyber-world-outsmarting-terrorists-hackers-and-spies> [Accessed 22 September 2021].
- Nai Fovino, I; Barry, G; Chaudron, S; Coisel, I; Dewar, M; Junklewitz, H; Kambourakis, G; Kounelis, I; Mortara, B; Nordvik, J.p; Sanchez, I. (Eds.) Baldini, G; Barrero, J; Coisel, I; Draper, G; Duch-Brown, N; Eulaerts, O; Geneiatakis D., Joanny G., Kerckhof S., Lewis A., Martin T., Nativi S., Neisse R., Papameletiou D., Ramos J., Reina, V; Ruzzante, G; Sportiello, L;

- Steri, G; Tirendi, S; 2020. *Cybersecurity, our digital anchor*, EUR 30276 EN, Publications Office of the European Union, Luxembourg,
- National Crime Agency, 2021. *Money laundering and illicit finance*. [Online] Available at: [money-laundering-and-illicit-finance](#) [Accessed 30 November 2021].
- National Crime Agency, 2021. *National Strategic Assessment of Serious and Organised Crime*. [Online] Available at: <https://www.nationalcrimeagency.gov.uk/who-we-are/publications/533-national-strategic-assessment-of-serious-and-organised-crime-2021/file> [Accessed 30 November 2021].
- Northern Illinois University, 2021 *Responsible Conduct in Data Management – Data Ownership* [Online] Available at: [Data Ownership \(hhs.gov\)](#) [Accessed 30 November 2021].
- Office of Director of National Intelligence, 2021. *Transnational organised crime - a threat to international security - manifests itself in various regions in different ways*. [Online] Available at: https://www.dni.gov/files/documents/NIC_toc_foldout.pdf [Accessed 30 November 2021].
- Rodgers, C., 2021. *CISO Insight: Every AWS Service is a Security Service*. [Online] Available at: <https://aws.amazon.com/blogs/enterprise-strategy/ciso-insight-every-aws-service-is-a-security-service/> [Accessed 22 September 2021].
- Scofield, M; 1998. *Issues of Data Ownership* [Online] Available at: [sPrint Version - Issues of Data Ownership | Business intelligence, data warehousing and analytics editorial from DMReview](#) [Accessed 30 November 2021].
- Scott, P. D., 2004. *Drugs, Oil, and war*. London: Rowman & Littlefield.
- Spannaus, A., 2012. *Interview: Antonio Costa - Former UNODC Head talks about Drugs and the World Banking System*. [Online] Available at: https://larouchepub.com/other/2012/3917costa_drugs_banks.html [Accessed 30 November 2021].
- Srnicek, N., 2017. *Platform Capitalism*. Polity Press : Cambridge.
- Statista Research Department, 2021. *Facebook: number of monthly active users worldwide 2008-2021*. [Online] Available at: <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/> [Accessed 27 September 2021].
- Statista, 2021. *Worldwide Amazon Users*. [Online] Available at: <https://www.statista.com/statistics/829113/number-of-paying-amazon-prime-members/> [Accessed 22 September 2021].
- The Economic Times, 2021. *Definition of Monopoly*. [Online] Available at: <https://economictimes.indiatimes.com/definition/monopoly> [Accessed 22 September 2021].
- The Free Dictionary by Farlex, 2021. *Anti-trust Law*. [Online] Available at: <https://legal-dictionary.thefreedictionary.com/antitrust+law> [Accessed 22 September 2021].
- Trading Economics, 2021. *North Korea GDP*. [Online] Available at: <https://tradingeconomics.com/north-korea/gdp> [Accessed 22 September 2021].
- Vailshery, L., 2021. *Apple's revenue worldwide 2004-2021*. [Online] Available at: <https://www.statista.com/statistics/265125/total-net-sales-of-apple-since-2004/> [Accessed 30 November 2021].
- Wang, B. & Li, F., 1995. *Information Warfare*. [Online] Available at: https://irp.fas.org/world/china/docs/iw_wang.htm [Accessed 22 September 2021].
- World Bank, 2021. *World Development Indicators*. [Online] Available at: <https://databank.worldbank.org/data/download/GDP.pdf> [Accessed 22 September 2021].
- Zuboff, S., 2019. *The Age of Surveillance Capitalism: The fight for a human future at the new frontier of power*. 1st ed. London: Profile Books.

Data Mining for the Security of Cyber Physical Systems Using Deep-Learning Methods

Bhagawan Nath^{1,2}, Timo Hamalainen¹ and Soundararajan Ezekiel³

¹University of Jyväskylä, Finland

²Ton Duc Thang University, HCMC, Vietnam

³Indiana University of Pennsylvania, PA, USA

bnath@student.jyu.fi

timo.t.hamalainen@jyu.fi

sezekiel@iup.edu

Abstract: Cyber Physical Systems (CPSs) have become widely popular in recent years, and their applicability have been growing exponentially. A CPS is an advanced system that incorporates a computation unit along with a hardware unit, allowing for computing processes to interact with the physical world. However, this increased usage has also led to the security concerns in them, as they allow potential attack vendors to exploit the possibilities of committing misconduct for their own benefit. It is of paramount importance that these systems have comprehensive security mechanisms to mitigate these security threats. A typical attack vector for a CPS is malicious data supplied by compromised sensors that are part of the CPSs. To combat this attack vector, many systems are secured through fault tolerance, including methods such as checkpointing to recover the system. Looking at the diverse nature of attacks and their ever growing complexities, traditional security approaches may not counter them efficiently, which creates a vacuum to be filled with sophisticated state-of-the-art techniques. In this paper, Deep Learning methods such as autoencoders, and Support Vector Machines are proposed to secure CPSs against these attacks. The networks in these applied methods are trained with a normal data profile devoid of any malicious data. Data collected from the system's sensors at specified intervals is used to form a data series and input to the neural networks. The networks compare and analyze new data to the normal profile to detect anomalies, if there is any. In the presence of anomalous data, the networks generate corrective action(s) for these sensors and the physical states they are recording. Through detection of anomalies, effective security of CPSs may be improved in addition to providing protection for the sensors. Moreover, the proposed method of securing CPSs opens up the possibility of further research by showcasing the applicability of neural networks in securing CPSs.

Keywords: Cyber Physical System, autoencoder, Support Vector Machine, Fault tolerance, Sensor Data, Cyber Attack

1. Introduction

Cyber Physical Systems (CPSs) have provided a way for communication and computing processes to interact through sensors and actuators with the physical world. As technology has progressed over the years and all systems have become connected to computer networks instead of being isolated and stand alone, thereby their functionality and interoperability needs to broaden the requirements for CPSs to interact with other systems in a more real-time fashion. As more services and applications have been added there has been an increase in the exposure area and hence exploitable vulnerabilities. The interactions a CPS has with the physical world are what makes this system more vulnerable to attacks that go beyond conventional cyber-attacks (Cardenas et al, 2008), (Meshram et al, 2017). It has been shown that the operation of a car can be disrupted and even disabled using simple methods (Checkoway et al, 2011). Another study shows that a Denial of Service attack can compromise the Controller Area Network bus and its functions on a system (Cho et al, 2016). A CPS can also be compromised through attacks on the sensors. By compromising the physical environment that the sensor reads, input can be generated that opens the system to attack. Other sensors that can be targeted by this are cameras, Light Detection and Ranging (LiDAR), and Global Positioning Systems (GPS) and can be on systems ranging from autonomous vehicles to ships.

The threat of these attacks has prompted many into finding ways to improve the attack resilience of CPSs, focusing specifically on the sensors, actuators, and communication components. Efficient solutions are typically achieved by developing a method that can approximate the normal states of the system accurately, allowing for continued control even with compromised components. This is advantageous, as both the error correction and the normal operations run from the same controller. This method has led into the creation of CPS checkpointing and recovery.

One way a Cyber Physical System can retain its integrity is through fault tolerance. In this process, the distinction between the physical and cyber physical worlds are utilized to provide seamless instruction, even upon system

failure. This is achieved through roll-back – roll-forward schemes. Two operations take place in this scheme, the roll-back recovery focuses on the cyber state while the roll-forward recovery focuses on the physical state. The cyber states are composed of computational information, such as data values, and the physical states regard the physical information available, such as a light sensor detecting light. Roll-back refers to the CPS accessing the previous checkpoint, where it can then circumvent the error. Roll-forward recovery refers to keeping the system rolling at current time. This method essentially handles failed estimated states caused by either attacks or failures.

This study is based on a study by Njilla et al, 2019, titled “Internet of Things Anomaly Detection using Machine Learning”, which investigates the application of DL methods (autoencoders) to detect anomalies. This study implements multiple DL models, autoencoder and one-class-SVM, specifically on CPS datasets and these models are tested to find the efficacy towards the stated problem.

The remainder of this paper is organized as follows: Section II describes the technical background of the techniques relevant in this study. Section III summarizes the methodology of securing CPSs with DL techniques to do such, that are used in this study. Section IV shows the results of our study and displays qualitative data to support these results. Section V discusses the impact of our study including the viability and effectiveness while exploring the future direction of our work.

2. Technical Background

2.1 Cyber Physical Systems

A Cyber Physical System is the improvement upon an embedded system, where a mechanism is either controlled or monitored by the system. These mechanisms can include autonomous vehicles, smart grids, robotic systems, and more. Rather than having just a single system controlling the machine, systems can be combined to create a system of systems. As demonstrated in Figure-1, these systems can share information, hardware, networks, and physical space to function more efficiently. In such a system, the details of its deployment must be clear during its creation, rather than on the go. This can be likened to cars at an intersection. In normal instances, the vehicles would communicate to each other in coming way to indicate who is going where. With a CPS, a system would emerge while the cars interact with each other to provide functionality but would not exist before or afterwards. This makes it so that the two vehicles would communicate to the intersection, rather than to each other, eliminating the need for hardware compatibility (Mosterman et al, 2016).

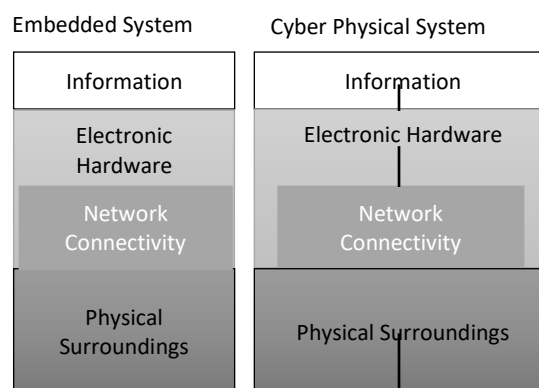


Figure-1: Embedded System vs a Cyber Physical System

There are numerous ways to model a CPS, with one of them being the 5-Layer model, shown in Figure-2, clearly demonstrates the architecture. At the base is connection, as acquiring data from the machines and components is the first step in a CPS. The correct sensors and equipment must be selected to give the system the most accurate data. Above that is the conversion layer, as the information gained through the connections must be translated into meaningful information. This allows for the machines in the CPS to have a form of self-awareness. At the centre is the cyber level, which acts as an information hub. Here, all the data is collected from the connected machines and analysed, allowing for the CPS to compare the performance of one machine to the rest in the system. The cognition level uses the previously obtained information and comparisons to make decisions on task prioritization. At the top is configuration, where feedback is given from the cyber side to physical side. Here the decisions made in the cognition level are passed to the system and monitored (Lee et al, 2015).

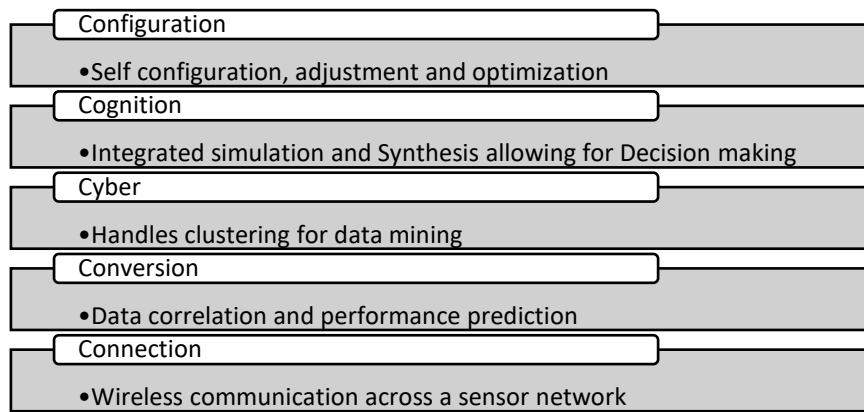


Figure-2: Five Layer Cyber Physical System Architecture

2.2 Generative Adversarial Networks

A Generative Adversarial Network (GAN) is an unsupervised deep learning technique that operates with two parts, the generative and the discriminative networks. These two models compete with each other in a form of a game setting. The GAN model would be trained on both - real data and fake data (generated by the generator). The discriminator's job is to determine fake data from real ones. The generator is a learning model, so starting from very basic at its initial stage, it is likely to produce low or even completely noisy data that does not reflect the real distribution or the properties of the real data.

The primary goal of the generator is generating artificial data that can pass the discriminator successfully. The model starts taking some noise, usually trying to convince the discriminator that the data produced by the generator is a legitimate one. The generator must learn how to trick the discriminator and win a positive classification (producing data classified as real). The generator's defeat is computed whenever any of those generated data is detected successfully as "fake". The discriminator, in turn, has to learn how to identify those fake data progressively. The discriminator is a loser whenever the model fails to recognize a fake data. The key concept is the simultaneous training of the generator and the discriminator. Hence, enhancing the overall performance of the system.

2.3 Deep convolutional neural networks

Deep convolutional neural networks (DCNNs) utilize deep, feed-forward architectures to learn the most significant features of their input. The trained model can then be used to classify new input into labels based on the learned features of the training data. In this regard, DCNNs are specific to the data on which they were trained, and new data must be able to be classified into existing labels (Staar et al, 2019). The feed-forward architecture consists of a varying number of layers stacked onto one another, using the previous layer's output as input for the next connected layer (LeCun et al, 2004), shown in Figure-3. The model is first initialized with all filters, weights, and parameters set to random values. Training data is then input into the network, which goes through a forward propagation step starting in the convolutional layer, which is comprised of filters that calculate an activation map containing the output of each convolution over the entire input (LeCun et al, 2004), (Lee et al, 2009). By activating when specific details are detected, the filters are used to learn the features of the input. The pooling layers of the network map the locations of the features of the data in relation to other features. The fully connected layers of the network then receive the activation mappings of the previous layers that contain the learned features, which are used for high-level reasoning and classification. At the output layer, the summation of the error is then calculated across all the classes and a backpropagation step is then applied using gradient descent to update the filter values and weights. The parameters of the model are subsequently optimized to improve classification accuracy.

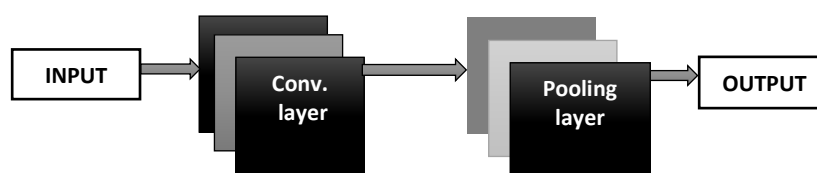


Figure-3: DCNN topology

2.4 Anomaly Detection

An anomaly is an out-of-way or abnormal behaviour shown by a system and anomaly detection is a data mining process that attempts to pinpoint data points, events or behaviours, that deviate from the normal behaviour of a dataset. Anomalies can be quite diverse and complex depending on various point of views, and transforming such an exceptional behaviour or an anomaly and its detection into an automated process can be even more challenging and time consuming task. The work starts with how we look at and define an anomaly, and it continues with the multitude of possible anomalies and different anomaly sources. An anomaly may be caused by unintentional or non-malicious causes such as communication errors, defective devices, noise in signals, significant environmental changes etc., which are usually unavoidable, often unpredictable and part of the system. But the other type of anomaly may be caused by intentional or malicious actions, such as a virus or another form of a malicious program, an intruder or a theorist. Overall, an anomaly detection method is usually picked based on a number of factors – different forms of anomalies, different anomaly sources, different types of systems (system behaviours) and different application domains (Sebestyen et al, 2017). Usually, an anomaly detection is used for data cleaning, fraud detection, intrusion detection, ecosystem instabilities, event detection in sensor networks etc.

It can be shown that the principal components are eigenvectors of the data covariance matrix. Therefore, Eigen decomposition of the data covariance matrix or singular value decomposition of the data matrix are often used to compute the principal components. PCA is the simplest of the true eigenvector-based multivariate analyses and it is closely related to factor analysis, which is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors (Njilla et al, 2019).

2.5 Autoencoders

Autoencoders are a type of learning model, that are used to create a lower dimensional representation of their input, through encoding. This allows for the most significant features of the input data to be learned by the model, with extraneous noise being filtered out, which can reduce high dimensional datasets without a significant loss of information. Autoencoders are comprised of two fully connected feedforward networks, an encoder and decoder, with a hidden layer between them, shown in Figure-4. Input data is first received by the encoder network, that extracts the most significant features and correlations from the data while reducing it into a compressed representation, in a latent space. The input then goes through a forward propagation step, through the layers of the encoder, with a variable number of nodes per layer. The hidden layer between the encoder and the decoder, obfuscates the data as to ensure that, the decoder does not duplicate the input. In general, the smaller the size of the hidden layer, the more compressed the encoded representation is. The decoder, which is an inverse architecture of the encoder receives the output of the hidden layer, and attempts to reconstruct the original data from the compressed representation. A loss function is then applied to measure the error, between the original and reconstructed data, and used to optimize the model. Autoencoders are specific to the data on which they are trained and can be used for tasks such as dimensionality reduction, denoising, and anomaly detection. In the case of multivariate data, for example, an autoencoder model can be trained to learn the normal conditions of a dataset. When new, previously unseen anomalous data is introduced to the network, it will encode, and attempt to reconstruct the original input. As the anomalous data contains data that differs from the normal learned conditions, it would have increased reconstruction loss.

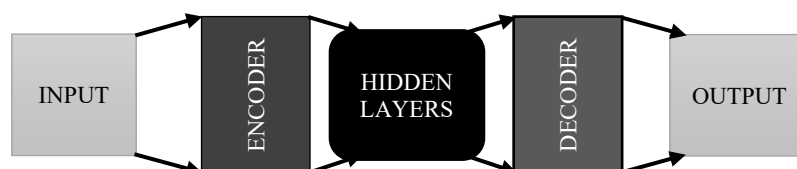


Figure-4: Autoencoder procedure

2.6 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. The goal of PCA is to extract the most significant information from the data, compress the dataset by keeping only this most significant information, and analyze the structure of the data observations and the variables (Markopoulos et al, 2014). To achieve these goals, PCA calculates new

variables called principal components which are obtained as linear combinations of the original variables. The first principal component is required to have the largest possible variance or inertia and therefore this component will “explain” or “extract” the largest part of the inertia of the data. The second component is calculated under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed likewise. The values of these new variables for the observations are called factor scores. The factor scores can be interpreted geometrically as the projections of the observations onto the principal components.

2.7 Data preprocessing

Data preprocessing is a very crucial step in any data mining process, as it determines the quality and relevance of meaningful insights from the data. It is due to the fact that the real-life data usually are incomplete, inaccurate, inconsistent and often lacks specific attribute values/trends (missing values). Thus, a careful screening of the data is must for such problems in order to produce clean, well formatted and organized ready-to-use raw data for the machine learning processes (Oliveri et al, 2019). The figure-5 demonstrates the application of data preprocessing needed prior to using them in a ML process.

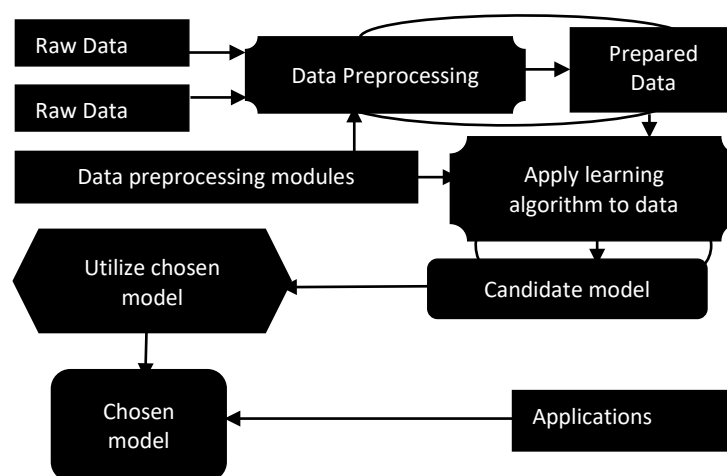


Figure-5: Data preprocessing

Data preprocessing starts with acquiring the data, cleaning the data, selection of instances, normalization, transformation, encoding the categorical data, splitting of datasets, feature scaling etc. Data preprocessing needs a careful insight as it may affect the way the outcomes of the final data processing can be interpreted (Oliveri et al, 2019), (Chatzigiannakis et al, 2006), (Simone et al, 2021).

2.8 Data mining

Data mining refers to the various techniques employed to detect correlations, anomalies, or patterns within the data so that insights about the data can be made. The mining does not refer to extraction of the data itself, but rather to the knowledge about the data that can be obtained. For this reason, an alternative term used in place of data mining is Knowledge Discovery in Databases (KDD). The KDD process consists of seven steps, shown in Figure-6. First is data cleaning, where missing values and noisy data are removed. This can be done through data discrepancy detection and data transformation tools. Next is data integration, where heterogeneous data from multiple sources are combined into a common source. This can be done through data migration or synchronization tools or through an extract-load-transformation process. Data selection comes next, deciding which data is relevant to the analysis and retrieving it from the data collection. Neural networks, decision trees, naïve bayes, clustering, and regression can be used to perform this selection. The fourth process is data transformation, converting the data into an appropriate form for the mining process using data mapping and code generation.

Next step is data mining, which is a process to identify interesting patterns and knowledge from a large amount of data. Pattern evaluation follows next, and it involves identifying interesting patterns representing some form of knowledge. Data summarization and visualization methods are used to convert the data to a form that users can understand. Finally comes Knowledge representation step, this is where data visualization and knowledge representation tools can be used to represent the mined data for a better and clearer understanding.

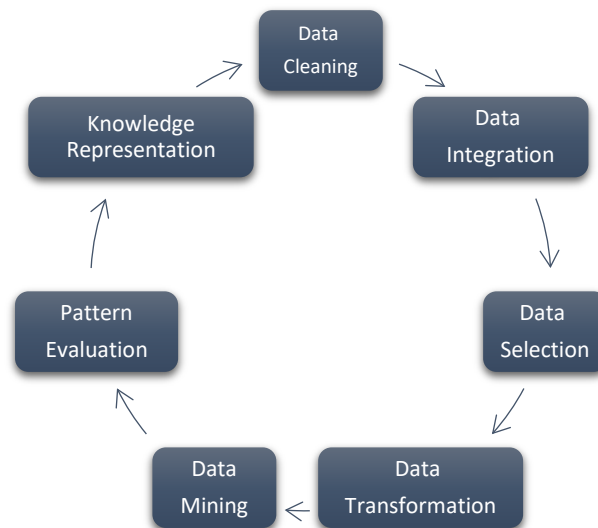


Figure-6: Data mining (KDD)

3. Methodology

The study is primarily based on two CPS datasets, the first one being a Water Distribution Testbed (WDT) dataset, which have been acquired from a water distribution hardware-in-the-loop testbed which emulates water passage between nine tanks via solenoid-valves, pumps, pressure and flow sensors, and the dataset is broadly divided into physical and network data. The other dataset, on the other hand, is a Supervisory Control And Data Acquisition (SCADA) Gas dataset.

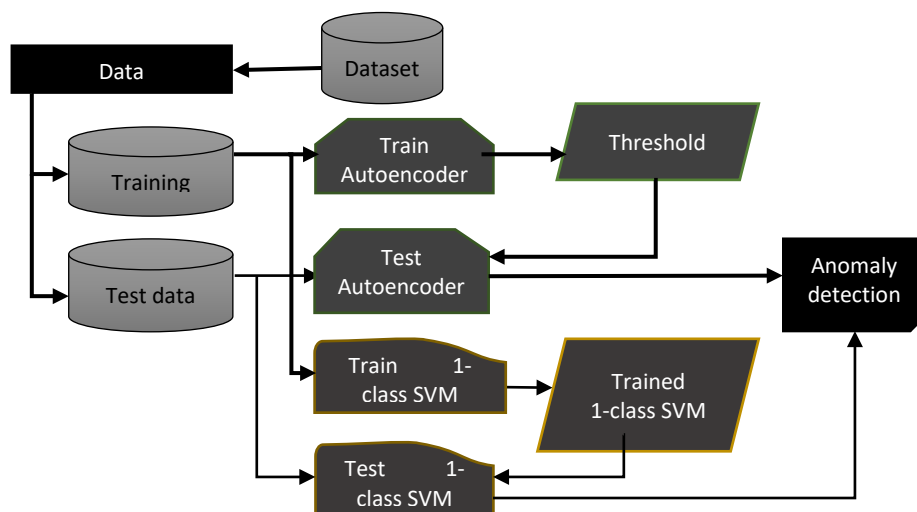


Figure-7: Basic methodology

The basic methodology is depicted in the figure-7, which starts with collecting CPS datasets from different sources, and then preprocessing them for a better study. The datasets are firstly preprocessed using one-hot encoding to represent certain categorical variables as binary vectors to ease the performance of the study, and the datasets are then normalized using Min-Max scaler. These datasets are individually divided into two parts – train and test data, out of which train data is picked as only normal data (without any anomalies) and test data containing both normal as well as anomalous data. The train data is further split into train and validation, certain amount of the train data is reserved to validate the models prior to doing the testing. Autoencoder and one-class SVM models are used in this study, they are trained using non-anomalous normal training data, validated with another set of data (validation data) and both the models are fine-tuned, wherever possible, for a better performance. For instance, in case of the autoencoder model, the parameters contributing to the reconstruction loss are determined and adjusted to minimize the loss. The one-class SVM is adjusted with its hyperplane parameters like kernel, gamma and nu, to achieve a satisfactory result. The models thus obtained are used to predict the anomalies from the test data and their performances are checked and compared individually.

4. Results

In case of autoencoder, different specifications of the important hyper-parameters such as depth of encoder-decoder, no. of nodes per layer in both, bottleneck etc. are tried and chosen to best suit the model. The model is later trained for 100 epochs with a batch size of 64, helping the encoder part of the model to compress and encode the parameters to a lower dimensional space. The latent space thus obtained is used by the decoder of the model to reconstruct the input data. The threshold is computed from the distribution of the training loss, which plays the key role in determining whether an observation is normal or anomalous. In case of one-class SVM, the upper bound on the fraction of outliers is determined carefully and set, the stopping tolerance that affects the number of iterations used while optimizing the model, and depends on the stopping criterion value, is also well thought and set, to enhance the performance of the model at an optimum level.

Looking at the purpose of the study being determining whether an observation is normal or anomalous, which is a classification problem, the performances of both the models are compared using the confusion matrix. It is a mechanism that not only provides accuracy, but also helps in estimating correct classification and misclassification. From the confusion matrix, important metrics such as accuracy, sensitivity, specificity, and precision are calculated using the following equations:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision}$$

Where, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

Sensitivity or recall and precision are critical class-level metrics, maximizing one can pull the value of the other one down and vice-versa, as they are inversely proportional to each other. In most of the problems, we could either prioritize maximizing either sensitivity or precision, depending on the nature of the problem. Thus, it demands a trade-off between sensitivity or recall and precision, and this is where the F1-score may be handy, which is the harmonic mean of these two metrics. A model is considered perfect if its F1-score is 1, and 0 indicates the model as a total-failure.

Table-A: Performance metrics of the one-class SVM model

One class SVM				
Accuracy	Sensitivity	Specificity	Precision	F1-score
94.09	94.05	95.56	99.87	96.87
95.19	95.22	92.43	99.93	97.52
97.03	96.95	97.62	99.69	98.30
96.60	96.63	91.87	99.94	98.26
96.60	96.64	94.36	99.88	98.24
95.90	95.90	94.37	99.86	97.84

Table-B: Performance metrics of the autoencoder model

Autoencoder				
Accuracy	Sensitivity	Specificity	Precision	F1-score
99.58	99.87	85.88	99.70	99.78
98.77	99.72	85.93	98.97	99.34
99.82	99.87	99.76	99.76	99.82
Autoencoder (contd.)				
Accuracy	Sensitivity	Specificity	Precision	F1-score
99.84	99.93	99.75	99.74	99.84
99.82	99.94	99.70	99.69	99.82
99.56	99.87	94.20	99.57	99.72

The performance metrics of both autoencoder and one-class SVM models are computed, and the vital metrics are represented in Table-A and Table-B in tabular format. In both the models, the performance metrics are computed multiple times to determine the consistency and also to rule out any foul play which may occur.

5. Conclusion and Future work

The results thus generated by both the models, with the datasets under consideration, indicate that the models perform well if they are individually fine-tuned, however it requires a considerable amount of time to find out the parameters and their efficient settings. While the two models are compared to each other, it is found that the autoencoder model performance is relatively better compared to that of the one-class SVM. However, the study doesn't generalize it as the scenario may be very different while using diverse datasets and the models may require a different tuning of the relevant parameters.

In future, this method may be applied into a constantly running CPS, allowing for the performance of corrective action(s) to be taken once an anomaly is found, which is beyond the scope of the current study.

References

- Cardenas, A. A., Amin S., and Sastry, S. (2008) Secure control: Towards survivable cyber-physical systems. In International Conference on Distributed Computing Systems Workshops (ICDCSW). IEEE.
- Chatzigiannakis, V. et al, (2006) Hierarchical Anomaly Detection in Distributed Large-Scale Sensor Networks, 11th IEEE Symposium on Computers and Communications (ISCC'06), pp 761-767.
- Checkoway, S. et al. (2011) Comprehensive Experimental Analyses of Automotive Attack Surfaces. In USENIX Security Symposium. San Francisco.
- Cho, K.-T. and Shin, G. K. (2016) Error Handling of In-vehicle Networks Makes Them Vulnerable. In ACM Conference on Computer and Communications Security (CCS).
- LeCun, Y, Huang, FJ and Bottou, L. (2004) 'Learning methods for generic object recognition with invariance to pose and lighting', Proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II97-II104.
- Lee, H. et al. (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp 609–616.
- Lee, J., Bagheri, B., and Kao, H.-A. (2015) A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems, *Manufacturing Letters*, Volume 3, ISSN 2213-8463, pp 18-23
- Markopoulos, Panos P., Karystinos, George N. and Pados, Dimitris A. (2014). Optimal Algorithms for L1-subspace Signal Processing. *IEEE*, Vol 62, pp 5046-5058.
- Meshram, A., Haas, C. (2017) Anomaly Detection in Industrial Networks using Machine Learning: A Roadmap. In: Beyerer J., Niggemann O., Kühnert C. (eds) *Machine Learning for Cyber Physical Systems. Technologien für die intelligente Automation (Technologies for Intelligent Automation)*. Springer Vieweg, Berlin, Heidelberg.
- Mosterman, P.J., and Zander, J. (2016) Industry 4.0 as a Cyber-Physical System study. *Softw Syst Model* 15, pp 17–29.
- Njilla, L. et al, (2019) Internet of Things Anomaly Detection using Machine Learning, 2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp 1-6.
- Oliveri, P. et al, (2019). The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial. *Analytica Chimica Acta*. 1058: pp 9–17.
- Sebestyen, G. and Hangan, A. (2017) Anomaly detection techniques in cyber-physical systems, *Acta Universitatis Sapientiae, Informatica*, vol.9, no.2, pp 101-118.
- Simone, G. et al (2021). A hardware-in-the-loop water distribution testbed (WDT) dataset for cyber-physical security testing. *IEEE Dataport*
- Staar B., Lütjen M., Freitag M. (2019) Anomaly detection with convolutional neural networks for industrial surface inspection, *Procedia CIRP*, vol. 79, pp 484-489

‘Out Beyond Jointery’: Developing a Model for Gaming Multi-Domain Warfare

Keith Scott

De Montfort University, Leicester, UK

jklscott@dmu.ac.uk

Abstract:

Play is battle and battle is play. [Huizinga, Homo Ludens]

What Huizinga is saying here is not that conflict is *playful*, but rather, it is a *game*, following set rules of conduct and occurs within a defined zone of action. Elsewhere in *Homo ludens*, he argues that modern warfare operates without the ritualised, rule-based structure of, for example, the mediaeval tourney. The purpose of this paper is to consider the ways in which a model based on the structure of games may help us better engage with the challenges of Multi-Domain Conflict. We are all familiar with the concept of Cyber as the 5th Domain of warfare, but we need to consider it not as a discrete zone, but as running through and interpenetrating the other 4 (Earth, Sea, Air, Space), the informational spine that enables all other forms of conflict. This paper will: 1. Discuss the developing concept of Multi-Domain Conflict as a move ‘beyond jointery’ (as General Sir Nick Carter put it) into a truly integrated form of warfare, blurring and collapsing boundaries between kinetic and non-kinetic, between the services, and between military and civilian authority; 2. Outline a theoretical model for conceptualising Multi-Domain Conflict as gamelike in form, with environments of operation (‘boards’), protagonists (‘players’), and possible forms of action (‘moves’). As befits a conference on Cyber and Information Warfare, it will argue that the D5 model of IW (Deny, Disrupt, Degrade, Deceive and Destroy) is portable and scalable across the other 4 domains (Land, Sea, Air, Space); 3. Show how this theoretical model can be employed both to model and simulate Multi-Domain Conflict; wargames have been a key element of military planning and training for at least a century – this paper argues that we need to develop a new *Kriegspiel* to better understand coming conflicts.

[...] theories are only made to die in the war of time. Like military units, they must be sent into battle at the right moment; and whatever their merits or insufficiencies, they can only be used if they are on hand when they’re needed. [Debord 1978, p.10]

Keywords: Multi-Domain Conflict; Serious Gaming; Information Warfare; Cyber Warfare

1. Introduction

In 2013, Thomas Rid published a work which claims categorically that *Cyber War Will Not Take Place*. This declaration is, I argue, fundamentally wrong, based as it is on a definition of ‘war’ which, as will be shown, is inadequate for understanding the complex and proteiform nature of present and future conflict. It is not simply the case that cyber warfare *does* exist as a discrete and developing form of warfare, where cyber ‘munitions’ blur the traditional division between kinetic and non-kinetic effects (Stone 2013; Farwell & Rozhinski 2011; Collins & McCombie 2012; Zetter 2014), but that conflict within the Cyber domain is an inevitable and necessary element of warfare. This paper will present three themes related to contemporary and future warfare:

1. an analysis of ‘war’ (arguably a highly contested term) in its actuality, moving away from the traditional Clausewitzian model towards a conception of a much broader and complex form of interlinked and interacting multi-domain conflict;
2. an attempt to develop an initial conceptual model for mapping, modelling and analysing multi-domain conflict;
3. a discussion of how such a conceptual model may form the basis for future wargame-based modelling and simulation exercises for all stakeholders (not merely the military) in an era of ‘unrestricted warfare’.

We begin with an apparently straightforward question: what *is* War? Thomas Rid argues that, within his definition of the term, ‘cyber warfare’ is not a true concept, as it does not automatically imply the use of *force*:

An act of war is always instrumental, and to be instrumental there has to be a means and an end: physical violence or the threat of force is the means; forcing the enemy to accept the offender’s will is the end.

[Rid: 2]

While he accepts that cyber actions can cause damage equivalent to that of kinetic weapons, he states that they do not cause physical harm on a scale equivalent to that of conventional ordinance’s ‘shock and awe’:

The use of cyber weapons that could inflict damage and pain comparable to the pummeling of Dresden, London, Belgrade, or Beirut at the receiving end of devastating airpower is, at present, too unrealistic even for a bad science fiction plot. [19]

Thirdly, he argues that war always ultimately requires a declaration of hostilities, and cannot be fought *sub rosa*, because 'History does not know of acts of war without eventual attribution' [2].

Ultimately, he contends that the very concept of 'cyber war' is an impossibility:

cyber war has never happened in the past, it does not occur in the present, and it is highly unlikely that it will disturb our future. [xiv]

Rid's argument is clear, meticulously structured and well-informed, and to a certain extent true, but only within the frame of reference he uses, which rests on a very narrow definition of what 'war' is. His model is essentially Clausewitzian, and this, I contend, is an inadequate model to understand conflict in our era, and indeed in considering conflict in general. Clausewitz died in 1831, in an era which was overwhelmingly pre-Industrial, where limitations of military technology (single-shot muskets, cannon with limited range and slow rate of fire, poorly armoured and slow cavalry...) restricted combat to overwhelmingly large-scale, close-range massed set-piece battles, where the objective was to destroy the enemy's forces and conquer the physical terrain. Moreover, warfare in Clausewitz's era was limited to two domains (land and sea; air warfare truly begins in the 1850s with the use of balloons for reconnaissance and limited offensive action); we now operate in five (land, sea, air, space, and cyber (including both information space and the electromagnetic spectrum). To employ the terminology employed by Lind et al (1989) and Hammes (2006), Clausewitz was fighting 1st-Generation warfare; we are now facing 4th Generation conflict.

It should not be thought, however, that my objection to Rid's use of a Clausewitzian standpoint is due to the latter's position as a historical figure; indeed, a figure from a much earlier period in the history of the development of strategic theory seems to offer a set of ideas which map closely onto the modern era. Living about 2000 years before Clausewitz, Sun Tzu presents a model of warfare which runs explicitly counter to Rid's. Where Rid argues that attribution is central to war, Sun Tzu prefigures the doctrine of maskirovka when he declares 'All warfare is based on deception' (*The Art of War*, p.3). Where Rid states that the employment of overwhelming physical force is an essential element of warfare, Sun Tzu that victory can be achieved in many ways: 'to fight and conquer in all your battles is not supreme excellence; supreme excellence consists in breaking the enemy's resistance without fighting' (*The Art of War*, p.8).

My main concerns with Rid's contention are twofold. Firstly, he does not seem to see that cyber actions can have effects that are as physically and psychologically devastating as kinetic warfare. We know that the SCADA systems which control much of our critical national infrastructure are hugely vulnerable to cyber attack (Shaw 2004; Yadav and Paul 2019; Trend Micro 2019). Any IT-controlled system is potentially vulnerable to attack and disruption, with potentially catastrophic results; consider for example the likely outcome of a coordinated series of attacks on the air traffic control systems at major airports (Tamimi, Hahn, and Roy. 2020; Hird 2020). Add to this the challenges posed by attacks to the financial system, and the opportunities for Information Warfare offered by social media, and the possibilities of cyber-created 'shock and awe' are all too evident. Just because we have (apparently) not yet seen such a large-scale attack does not mean that it could not or will not happen, nor that we should not prepare for it. Thinking the unthinkable is a key element of effective security. We must also consider that a networked society, where all zones of human activity – civil and military, private and public sector, individual and collective – are intimately and inevitably interconnected leads to a vastly increased attack surface. One of the fundamental principles behind this paper is that conflict is now inherently *multi-domain*, and that actions in one domain can and must be evaluated for their possible effects in other domains. The cyber domain in effect forms an informational spine which runs through (and in many cases runs) actions in all the other domains, and effective cyber security must draw on the expert knowledge of those who work in all other spheres of human activity. Put crudely, 'what happens in cyber does not stay in cyber'.

My second objection to Rid's conception of warfare is more wide-ranging; it rests on a very simple question, the answer to which is highly complex: what *is* war? There is in fact no single definition in international law of the term; we have codes of conduct concerning behaviour during war (most notably the Geneva Conventions), but no official summation of what war actually consists of. If it is a large-scale open conflict between nation states,

then the 'War on Terror' is not a war (neither Al-Qaeda nor ISIS/ISIL/Daesh are nation states). If an official declaration of war is required, then the UK has not been at war since 1942, when it declared war on Siam; similarly, the last declaration of war by the United States was also in 1942, against Bulgaria, Hungary, and Romania. Hence, the 'wars' in Vietnam, Korea, the Falklands, and Iraq cannot be officially named as such; this is clearly ludicrous. Consider the wikipedia entry for 'Outline of war'; there is no single overarching definition, merely a quasi-fractal proliferation of subcategories. There are *wars* (asymmetric, network-centric, unconventional, psychological...) but no *war*.

Jason Rivera (2014) maps out a taxonomy of forms of conflict, marking a clear delineation between 'war' and 'not-war' (Figure 1 below); such a neat division is arguably illusory.

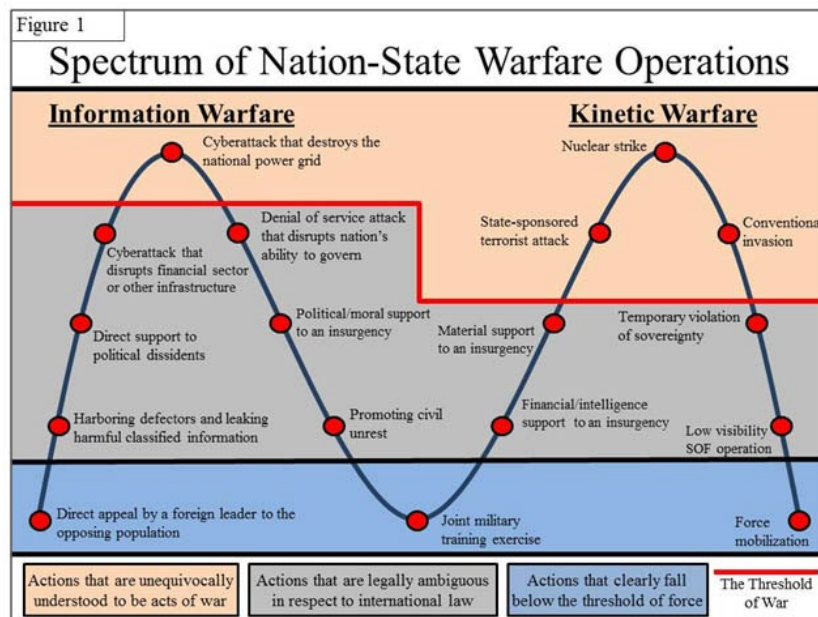


Figure 1: a mapping of 'war' and 'not-war' (Rivera 2014).

Examine a selection of the recent work on grey zone conflict (Carment and Belo 2020; Hoffman 2018; Morris *et al* 2019; Raine 2019 *inter alia*) and what becomes clear is that this hybrid, liminal form of 'warfare' is becoming ever-more common, posing a profound challenge to existing models of the war/peace division. As Rosa Brooks (2016) observes, it is perhaps time that we

accept that "war" and "peace" are not binary opposites, but rather the outer limits of a continuum. Indeed, add in cyber, individualized weapons, and various nonlethal forms of coercion and control and a two-dimensional continuum may not be enough: we may need to conceptualize warfare in three dimensions, or even more.

The conceptual basis for such conflict has been advanced by Qiao and Wang, in the work translated as *Unrestricted Warfare*. These two officers of the PLA argue that war can be fought in many ways, kinetic and non-kinetic, mentally and physically, across all spheres of activity:

all means will be in readiness, that information will be omnipresent, and the battlefield will be everywhere. It means that all weapons and technology can be superimposed at will, it means that all the boundaries lying between the two worlds of war and non-war, of military and non-military, will be totally destroyed

Coleman (2009; 2006) lays out what he terms the '15 modalities of Unrestricted Warfare' as

1. Cultural warfare
2. Economic warfare
3. Environmental warfare
4. Financial warfare
5. Illegal drugs warfare
6. International law warfare
7. Information and media warfare

8. Telecommunications and network warfare
9. Political warfare
10. Psychological warfare
11. Resource warfare
12. Smuggling warfare
13. Technology warfare
14. Terrorist warfare
15. Gang warfare

(he sees modalities 7, 8, and 13 as the ambit of Cyber warfare; 9 and 10 should surely also be considered as falling within its domain).

Rid, with his narrow neo-Clausewitzian conception of war, has no adequate way of responding to Qiao and Wang's infinitely wider and more open vision of how conflict can be conducted in many different ways in many different arenas. If we imagine combat between forces driven by these two opposed mindsets, it seems hard to see that Rid's side would emerge victorious. Qiao and Wang display a willingness to work across traditional barriers, and to embrace as wide a range as possible of strategies and tactics:

only if we break through the various kinds of boundaries in the models of our line of thought, take the various domains which are so completely affected by warfare and turn them into playing cards deftly shuffled in our skilled hands, and thus use beyond-limits strategy and tactics to combine all the resources of war, can there be the possibility that we will be confident of victory. [200-01]

War has always been a complex, proteiform entity; what changes is our belief as to what falls within the class of 'acceptable forms of war'. If, however, we do as Rid does, and seek to deny the existence of forms which lie beyond those of which we approve, then we will inevitably fail. The model of Unrestricted Warfare gives us the 'what' and 'how' of modern warfare; to understand the 'where' we must consider the growing interest in Multi-Domain warfare.

2. Multi-Domain Conflict

In 2019, the then Chief of the UK Defence Staff, General Sir Nick Carter, delivered a speech in which he made many of the same points outlined above, arguing that the traditional idea of 'war' and 'peace' as discrete states is illusory. He discussed certain (unnamed) nation states, and quoted Edward Lucas (2019), saying:

these regimes [...] view the strategic landscape as characterised by a continuous and never-ending struggle that encompasses everything from what the West calls 'peace' to nuclear war. When they consider conflict along this spectrum, the primary change from one end to the other is the relative weighting that is given to non-military and military instruments. These regimes believe that they are already engaged in an intense form of warfare, but it is political conflict and not kinetic warfare. Their primary operational focus is on employing a range of mainly non-military instruments in non-traditional ways below the threshold of large scale conventional military operations to achieve strategic gains [Carter 2019].

We should note that in Lucas' original article, he actually names the states involved; it should come as no surprise that one of them is the state whence comes the theory of unrestricted warfare. Carter and Lucas are making precisely the point made throughout this paper, that 'war' and 'peace' occupy opposite poles on a continuum, or a cycle:

The linear depiction of peace to war and back again must be revised to reflect the cyclical nature of war where there are only positions of relative advantage. The continuum of conflict is defined by competition short of conflict, conflict itself, and the return to competition (Perkins 2017)

This conception of war as cyclic, rather than linear, shifting constantly and endlessly from high- to low-intensity and back again, is matched by the appearance in recent years of a model of military operations which is truly holistic, and collapses the traditional divisions between different services' domains of activity; we are, it is

argued, in the age of ‘multi-domain warfare’. A review of some of the most recent work in this field¹ shows an attempt to, in Carter’s words, ‘move beyond jointery’ and develop fighting forces which can operate in a truly integrated manner, going far beyond the idea of ‘combined operations’ (the logical extension of the idea of multi-domain conflict, namely abandoning the tradition divisions between the various branches of the armed services, has not as yet truly been engaged with, for reasons it is not hard to determine). The UK Ministry of Defence’s *Joint Concept note 1/20* depicts the way in which the multi-domain model sees the 5 domains of conflict operating (Fig.2 below); through a series of overlapping zones of effect, Air wraps around Land and Sea, Space wraps around these, and the fifth domain (here defined as ‘Cyber and electromagnetic’) surrounds them all. This is a fair first step at conceptualising the multi-domain environment, but it is flawed, for 2 key reasons:

1. **Conceptual Incompleteness:** this model still suggests that the domains are discrete, if overlapping, zones, whereas a truly multi-domain conflict will operate through and within multiple domains simultaneously, and the boundaries between the zones are at best a matter of convention, rather than reality (for example, which domain does a UAV controlled by a land unit launching an attack from the air on a maritime target ‘belong’ to?). Secondly, and crucially, this model fails to grasp the full complexity of the relationship between the ‘cyber’ domain and all the others; it does not simply envelop the others, it runs through them, penetrating and connecting them, the informational skeleton or nervous system without which none of the others can operate. Any map, of course, is a reductive model of reality, a 2-D depiction of a 3-D environment, but multi-domain warfare requires a ‘map’ which can better present the interconnection and interpenetration of actions and effects (kinetic and non-kinetic).
2. **Limitation of Scope:** It is perhaps unsurprising that a model devised by the Ministry of Defence should consider multi-domain conflict as a purely military affair; however, it is inadequate for examining war as it truly is (and arguably always has been). Firstly, war has always affected civilian populations, either as victims caught in the crossfire, or as deliberate targets. Secondly, warfare in its ‘unrestricted’ form is ‘fought’, not just by soldiers in a combat environment, but non-kinetically, in the economic, political, and sociocultural domains. If we cannot include the non-military in our model, we fail to grasp the true nature of the situation, and our attempts to achieve true full-spectrum dominance are doomed to failure before we begin. Multi-domain conflict is by its essential nature multi-agency and multi-stakeholder, and to be successful, we must find a way of modelling such a hugely complex system of systems. As recent examples of Whole of Government and multinational operations during the pandemic have shown, co-ordinating action across the civil powers of a single nation state is challenging enough; to be truly successful, we need to be able to consider how to operate across the traditional borders between civil and military powers (without surrendering to totalitarianism and a one-party state) and national boundaries. What follows is an attempt to devise an initial mapping of such a system, designed to allow us to *analyse* multi-domain conflict, and to *conduct* it more effectively. As George Box (1976) said, ‘all models are wrong’ (p.792); the following model seeks to be slightly less wrong than those currently in operation.

¹ See, *inter alia*, Clare 2020; Gady and Stronell 2020; Jones and Diaz de Leon 2020; McCoy 2018; Ministry of Defence 2020; TRADOC 2017 and 2021.

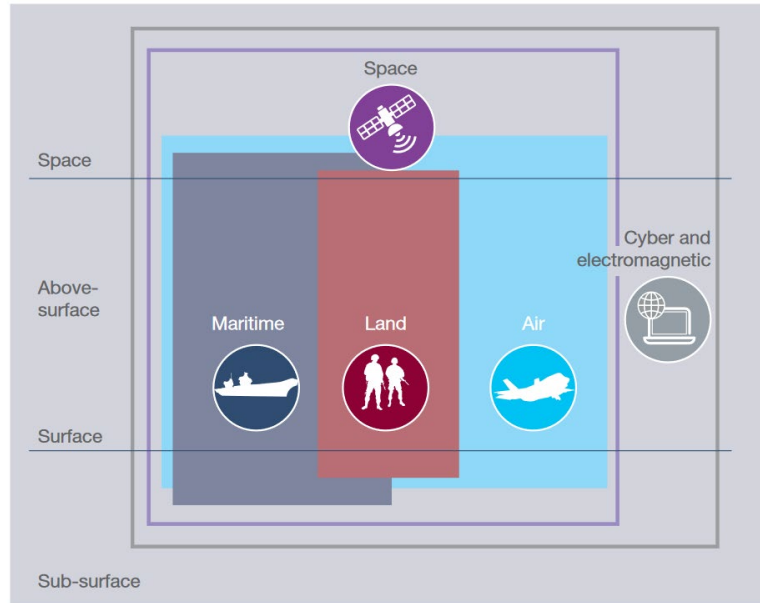


Figure 2: Illustration of domains as per UK doctrine (Ministry of Defence 2020: p. 18)

3. Modelling the Multi-Domain

The following model is adapted from a conceptual framework for understanding and designing Influence Operations (Kodalle *et al.* 2019). Any abstraction of reality is inevitably a simplification, and it would be easy to devise an ever-more detailed taxonomy of the battlespace; at this stage, we are concerned with building a model that *works* more or less effectively. If we accept that the forms of conflict we will encounter in the future will cut across the ‘war/peace’ boundary, then we need to be able to develop an abstraction that does the same. The basis of this model is a matrix formed of the intersection of the categories used in the PMESII and ASCOPE analytical techniques. The PMESII framework divides an operational environment into the following zones:

- P** – Political.
- M** – Military.
- E** – Economic.
- S** – Social.
- I** – Infrastructure.
- I** – Information.

This framework is then set alongside the complementary ASCOPE methodology, used to gain a better understanding of an Operational Environment in a COIN context. ASCOPE refers to:

A	–	Areas
S	–	Structure
C	–	Capabilities
O	–	Organizations
P	–	People
E – Events		

The PMESII/ASCOPE frameworks lead to creation of a 6x6 matrix, allowing us to define target areas as variables (PMESII) within the dimensions offered by ASCOPE (see Fig. 3 below).

	P Political	M Military	E Economic	S Social	I Information	I Infrastructure
A Areas	Areas - Political (District Boundary, Party affiliation areas)	Areas - Military (Coalition / LN bases, historic ambush/IED sites)	Areas - Economic (bazaars, shops, markets)	Areas - Social (parks and other meeting areas)	Areas - Information (Radio/TV/newspapers /where people gather for word-of-mouth)	Areas - Infrastructure (Irrigation networks, water tables, medical coverage)
S Structures	Structures - Political (town halls, government offices)	Structures - Military / Police (police HQ, Military HHQ locations)	Structures - Economic (banks, markets, storage facilities)	Structures - Social (Churches, restaurants, bars, etc.)	Structures - Information (Cell / Radio / TV towers, print shops)	Structures - Infrastructure (roads, bridges, power lines, walls, dams)
C Capabilities	Capabilities - Political (Dispute resolution, Insurgent capabilities)	Capabilities - Military (security posture, strengths and weaknesses)	Capabilities - Economic (access to banks, ability to withstand natural disasters)	Capabilities - Social (Strength of local & national ties)	Capabilities - Info (Literacy rate, availability of media / phone service)	Capabilities - Infrastructure (Ability to build / maintain roads, walls, dams)
O Organizations	Organizations - Political (Political parties and other power brokers, UN,)	Organizations - Military (What units of military, police, insurgent are present)	Organizations - Economic (Banks, large land holders, big businesses)	Organizations - Social (tribes, clans, families, youth groups, NGOs / IGOs)	Organizations - Info (NEWS groups, influential people who pass word)	Organizations - Infrastructure (Government ministries, construction companies)
P People	People - Political (Governors, councils, elders)	People - Military (Leaders from coalition, LN and insurgent forces)	People - Economic (Bankers, landholders, merchants)	People - Social (Religious leaders, influential families)	People - Info (Media owners, mullahs, heads of powerful families)	People - Infrastructure (Builders, contractors, development councils)
E Events	Events - Political (elections, council meetings)	Events - Military (lethal/nonlethal events, loss of leadership, operations, anniversaries)	Events - Economic (drought, harvest, business open/close)	Events - Social (holidays, weddings, religious days)	Events - Info (IO campaigns, project openings, CIVCAS events)	Events - Infrastructure (road / bridge construction, well digging, scheduled maintenance)

Figure 3: The PMESII/ASCOPE matrix²

The matrix is now expanded from two dimensions into three, adding a layer for each of the five domains of warfare (Land, Sea, Air, Space, Cyber). This then creates a cube composed of $6 \times 6 \times 5 = 180$ cells, allowing us a high degree of precision in defining the origin, target, and spread of an attack (fig. 4).

This basic model lends itself easily to tailoring and/or refinement. In its original form, the 3-D matrix was composed of a $6 \times 6 \times 4$ framework, corresponding to PMESII+ASCOPE+DIME (Diplomacy, Information, Military, Economy). A further level of detail was added by giving each of the component blocks one of three 'spins' (analogous to quarks) to correspond to the strategic, tactical, and operational levels. Each layer of the basic framework could also be used as the base for a new 3-D $5 \times 5 \times 5$ matrix, with the Z-axis representing the 5 ways of responding to an attack (or indeed the 5 forms an information attack may itself take): *disrupt, deny, degrade, destroy, or deceive*.

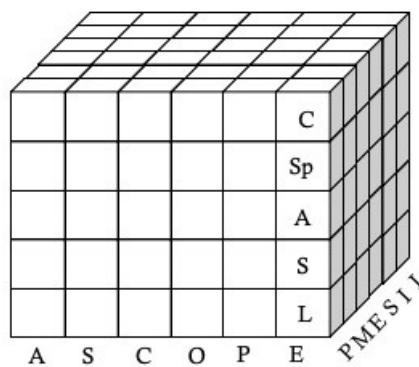


Figure 4: the 3-D model of the multi-domain battlespace, ASCOPExPMESIIx5 Domains of warfare (Land, Sea, Air, Space, and Cyber)

A 3-D model is of course much more challenging to visualise than a 2-D map, but the development of VR and AR technology offers a number of possible ways for employing this model; it gives us the ability to trace the path taken by an attack from one domain moving into others, and tracking the spread of its effect. (Consider a cyber

² Adopted from an open source planning template: <https://www.trngcmd.marines.mil/Portals/207/Docs/wtbn/MCCMOS/Planning%20Templates%20Oct%202017.pdf?ver=2017-10-19-131249-187>.

attack launched from a land base, transmitted through a satellite, to disable or degrade the effectiveness of maritime GPS systems; if such systems are carried on commercial freight shipping, this will effect international supply lines, and have a further effect on a national economy, civilian morale, and political stability.) Such a process of modelling and analysis can be enhanced by feeding in real-time data from the ever-growing range of sensors proliferating in civilian and military environments, while AI-driven Big Data analysis could see this form the basis for a predictive early warning system for future conflicts. Even at this initial stage, this model offers a means of analysing possible future events, and invites the user to think beyond the conventional idea of Force A launches Attack B on Target C, by presenting a much wider and deeper possible attack surface.

4. Next Steps

As stated throughout this paper, the model presented above makes no claims to be definitive; it does however seem to offer a more comprehensive and detailed conception of the battlespace within which multi-domain, unrestricted conflict can, does, and will operate. At a theoretical level, it seems to offer promise; how can we test it at a *practical* level? The obvious answer (assuming that starting a war is seen as undesirable), is to engage in a simulation or wargame. The matrix that forms the basis for this model was devised during the creation and testing of a wargame (Ormrod *et al* 2019) designed to simulate influence campaigns on an international level, taking a Whole of Government perspective and constructing teams of players from the military, governmental, business and academic sectors (precisely the multi-stakeholder perspective needed for combatting unrestricted warfare). In order to develop and refine the model proposed here, it must be tested, and the aim for future work is to scope the requirements for putting such an exercise into operation. As done in the 2019 exercise, the steps required will be:

1. assemble a team of interested stakeholders and subject matter experts to *design* the exercise, and a parallel group, drawn from as wide a pool as possible, to *test* it;
2. design an environment/gameworld within which the events of the exercise will occur;
3. devise a scenario (or set of scenarios) inviting/requiring a multi-domain response, with the use of our model as a support/guide for action and response;
4. playtesting (with an emphasis on Red Teaming; those engaged in defence must always consider the mindset of those *attacking* them – in order to protect a system, it is vital to know how it can be destroyed);
5. recording, reviewing, and revising the model, on the basis of the exercise.

We know that wargaming is a powerful tool for analysing previous actions and positing future events; the fact that the gaming of multi-domain conflict has already begun³ shows that it is seen as potentially helpful for considering the evolution of conflict. What this paper has sought to do is to challenge the ways in which we have previously conceptualised warfare, and to offer an initial step towards a better understanding and preparation for may lie ahead. The model it presents is portable and scalable, but it does not claim to be definitive.

References

- Box, G. E. (1976). "Science and Statistics", *Journal of the American Statistical Association*, Vol. 71, No. 356 (Dec., 1976), pp. 791-799.
- Brooks, R. 2016. *How Everything Became War and the Military Became Everything: Tales from the Pentagon*, Simon & Schuster, New York. Kindle Edition.
- Carment, D., and Belo, D. (2020) "Gray-zone Conflict Management: Theory, Evidence, and Challenges", *Journal of European, Middle Eastern, & African Affairs*, 9 June, [online], www.airuniversity.af.edu/JEMEAA/Display/Article/2213954/gray-zone-conflict-management-theory-evidence-and-challenges/.
- Carter, N. (2019) "Annual Chief of the Defence Staff Lecture and RUSI Christmas Party 2019", *gov.uk*, 5 December, [online], www.gov.uk/government/speeches/chief-of-the-defence-staff-general-sir-nick-carters-annual-rusi-speech.
- Clare, P. (2020). "The Answer is Multi Domain Operations – Now What's the Question?". *Wavell Room*, 13 February, [online], wavellroom.com/2020/02/13/the-answer-is-multi-domain-operations-now-whats-the-question/.
- Coleman, K. (2006). "The Challenge of Unrestricted Warfare - A Look Back and a Look Ahead", *Directions*, 12 January, [online], <https://www.directionsmag.com/article/3039>.
- Coleman, K. (2009). "Confronting UnRestricted Warfare", *military.com*, 3 March [online], <https://www.military.com/defensetech/2009/03/03/confronting-unrestricted-warfare>.
- Collins, S. and McCombie, S. (2012) "Stuxnet: the emergence of a new cyber weapon and its implications", *Journal of Policing, Intelligence and Counter Terrorism*, Vol.7, no. 12, pp.80-91.

³ Freedburg 2019; Jensen 2019; Lockheed Martin. 2019; Schuety and Hill 2018; Woodford 2019a and 2019b

- Debord, G., tr. Knabb, K. (1978). "In girum imus nocte et consumimur igni", the anarchistlibrary.org [etext].
- Farwell, J.P. and Rohozinski, R. (2011) "Stuxnet and the Future of Cyber War", *Survival*, 53:1, pp. 23-40.
- Freedburg, S.J. (2019). "Army Multi-Domain Wargame Reveals C2 Shortfalls", *Breakingdefense.com*, 12 September, [online], <https://breakingdefense.com/2019/09/army-multi-domain-wargame-reveals-c2-shortfalls/>.
- Gady, F-S, and Stronell, A. (2020). "Cyber Capabilities and Multi-Domain Operations in Future High-Intensity Warfare in 2030", NATO Cooperative Cyber Defence Centre of Excellence, Tallinn.
- Hammes, T.X. (2004) *The Sling and the Stone: On War in the 21st Century*, Zenith Press, St Paul, Mn.
- Hird, J. (2020) "AIR TRAFFIC MANAGEMENT: A CYBERSECURITY CHALLENGE", *Transport Security International*, 21 August 2020, [online], www.tsi-mag.com/air-traffic-management-a-cybersecurity-challenge/.
- Hoffman, F.G. (2018) "Examining Complex Forms of Conflict: Grey Zone and Hybrid Challenges", *Prism*, Vol.7, no.4, pp. 30-47. cc.o.ndu.edu/Portals/96/Documents/prism/prism7_4/181204_Hoffman_PDF.pdf?ver=2018-12-04-161237-307
- Jensen, B. (2019). "Welcome to Fight Club: Wargaming the Future", warontherocks.com, 4 January, [online], warontherocks.com/2019/01/welcome-to-fight-club-wargaming-the-future/.
- Jones, M.A. and Diaz de Leon J. (2020). "Multi-Domain Operations", *Three Swords Magazine*, November, p.38-41. jwc.nato.int/application/files/5616/0523/5418/issue36_08lr.pdf.
- Kodalle T, Ormrod, D., Sample, C., and Scott, K. (2020). "A General Theory of Influence in a DIME/PMESII/ASCOP/IRC2 Model", *Journal of Information Warfare*, Vol. 19, Issue 2, Spring, [online], <https://www.jinfowar.com/journal/volume-19-issue-2/general-theory-influence-dimepmesiiascopirc2-model>.
- Lind, W.S., Nightengale, K., Schmitt, J.F., Sutton, J.W., and Wilson, G.J. (1989). *The Changing Face of War: Into the Fourth Generation*, *Marine Corps Gazette*, October 1989, pp. 22-26.
- Lockheed Martin. (2019). "Learning through Wargaming: Helping Define the Roadmap for Multi-Domain Operations", Defense News, 13 September, [online], www.defensenews.com/native/lockheed-martin/2019/09/13/learning-through-wargaming-helping-define-the-roadmap-for-multi-domain-operations/.
- Lucas, E. (2019). "The West slumbers as its enemies run amok", *Standpoint*, 18 September, [online], standpointmag.co.uk/the-west-slumbers-as-its-enemies-run-amok/.
- Luke, B. (2012). *Recognizing And Adapting To Unrestricted Warfare Practices By China*. Graduate thesis, US Air War College, Montgomery, Al.
- McCoy, K. (2018). "Competition, Conflict, And Mental Models Of War: What You Need To Know About Multi-Domain Battle", *Modern War Institute*, 26 January, [online], mwi.usma.edu/competition-conflict-mental-models-war-need-know-multi-domain-battle/.
- Ministry of Defence. (2020) *Joint Concept Note 1/20: Multi-Domain Integration*, Development, Concepts and Doctrine Centre, Shrivenham.
- Morris, L. J., Mazarr, M.J., Hornung, J. W., Pezard, S., Binnendijk, A., and Kepe, M. (2019) *Gaining Competitive Advantage in the Gray Zone: Response Options for Coercive Aggression Below the Threshold of Major War*, RAND, Santa Monica, CA.
- Ormrod, D., Scott, K., Scheinman, L., Kodalle, T. Sample, C. Turnbull, B., and Ormrod, A. (2019) "The Persuasion Game: Serious Gaming Information Warfare and Influence". *Journal of Information Warfare*, Vol. 19, Issue 2, Spring, [online], www.jinfowar.com/journal/volume-19-issue-2/persuasion-game-serious-gaming-information-warfare-influence.
- Perkins, D.G. (2017) "Multi-Domain Battle The Advent of Twenty-First Century War", *Military Review*, November-December, [online], www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/November-December-2017/Multi-Domain-Battle-The-Advent-of-Twenty-First-Century-War/.
- Qiao, L., and Wang, X. (2020) *Unrestricted Warfare: China's Master Plan To Destroy America*, Shadow Lawn Press, Lambertville. [Kindle Edition].
- Raine, J. (2019) "War or peace? Understanding the grey zone", *IIIS.org*, April 3, [online], www.iiis.org/blogs/analysis/2019/04/understanding-the-grey-zone.
- Reynolds, L. (2019) "ASPI International Conference: War in 2025", *Australian Government Department of Defence Ministers*, June 13, [online], <https://www.minister.defence.gov.au/minister/lreynolds/speeches/aspi-international-conference-war-2025>.
- Rivera, J. (2014). "Understanding and Countering Nation-State Use of Protracted Unconventional Warfare", *Small Wars Journal*, 25 October, [online], smallwarsjournal.com/jrnl/art/understanding-and-countering-nation-state-use-of-protracted-unconventional-warfare.
- Schuety, C., and Hill, L. (2018). "An Air Force 'Way Of Swarm': Using Wargaming And Artificial Intelligence To Train Drones", warontherocks.com, 21 September, [online], warontherocks.com/2018/09/an-air-force-way-of-swarm-using-wargaming-and-artificial-intelligence-to-train-drones.
- Shaw, W.T. (2004) SCADA System Vulnerabilities to Cyber Attack, *EE Online*, September/October, [online], electricenergyonline.com/energy/magazine/181/article/SCADA-System-Vulnerabilities-to-Cyber-Attack.htm.
- Stone, J. (2013) "Cyber War Will Take Place!", *The Journal of Strategic Studies*, Vol. 36, No. 1. pp. 101-108.
- Tamimi, A., Hahn, A., and Roy, S. (2020) "Cyber Threat Impact Analysis to Air Traffic Flows Through Dynamic Queue Networks", *ACM Transactions on Cyber-Physical Systems*, Vol. 4, Issue 3 (May), pp. 1-22.
- TRADOC. (2017) *Multi-Domain Battle: Evolution of Combined Arms for the 21st Century 2025-2040*, U.S. Army Training and Doctrine Command (TRADOC), Fort Eustis, VA.
- TRADOC. (2021) *The U.S. Army in Multi-Domain Operations 2028*, U.S. Army Training and Doctrine Command (TRADOC), Fort Eustis, VA.

- Trend Micro. (2019) "One Flaw too Many: Vulnerabilities in SCADA Systems", *Trend Micro*, 16 December, [online], www.trendmicro.com/vinfo/us/security/news/vulnerabilities-and-exploits/one-flaw-too-many-vulnerabilities-in-scada-systems.
- Tzu, Sun, tr. Lionel Giles. (2000) *The Art of War*, Allendale Online, Leicester.
- Woodford, S. (2019a). "Wargaming Multi-Domain Battle: The Base Of Sand Problem", *Mystics and Statistics*, 11 January, [online], www.dupuyinstitute.org/blog/2019/01/11/wargaming-multi-domain-battle-the-base-of-sand-problem/.
- Woodford, S. (2019b). "What Multi-Domain Operations Wargames Are You Playing? [Updated]", *Mystics and Statistics*, 25 January, [online], www.dupuyinstitute.org/blog/2019/01/25/what-multi-domain-operations-wargames-are-you-playing/.
- Yadav, G. and Paul, K. (2019), "Assessment of SCADA System Vulnerabilities", *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1737-1744.
- Zetter, K. (2014) *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. Random House, New York.

Research Gaps and Opportunities for Secure Access Service Edge

Stephanus Petrus van der Walt and Venter Hein

University of Pretoria, South Africa

svdwalt@gmail.com

hventer@cs.up.ac.za

Abstract: This paper provides a contemporary discussion of security as a service from a network perspective and discusses state-of-the-art research conducted within a framework known as Secure Access Service Edge (SASE) (pronounced 'sassy'). SASE is a network security framework proposed by Gartner (2019) (MacDonald, Orans and Skorupa, 2019). This paper gives brief description of cloud concepts and technologies, focuses on network security in the cloud and aims to provide researchers with subjects of future research in SASE. To achieve the aim, the authors evaluate existing papers on SASE and its core components to identify gaps in the literature currently available on SASE.

Keywords: Secure access service edge, secure web gateway, zero trust network access, cloud access broker, digital forensic readiness

1. Introduction

Accessing the Internet is fraught with risk in the form of potential attacks and systems compromise. These risks and threats have forced organisations to make use of their own isolated networks that are typically isolated and protected by a single security stack at head office. With more applications and services having moved to the cloud and users working remotely (a situation that was accelerated by COVID-19), it is becoming impossible to isolate all network traffic. The vulnerability of the cloud became evident in cyberattacks on cloud services (ransomware, data theft, etc.) over the past two years and in the impact it had on organisations. Due to these vulnerabilities, it is imperative to protect all cloud-offered services in the same way that we currently protect our private networks.

The cloud offers virtualised computing, storage and networking resources over the Internet to users in a dynamic way. Owning cloud services is less expensive and much more elastic to scale than local servers. These advantages encourage customers to move their applications and services to the cloud, despite the security risk associated with such services. Traditional network security architectures were designed where the data centre was the crucial point of access for devices and users. With businesses' digital transformation and their adoption of Software as a Service (SaaS) and other cloud services, the traffic flow patterns have changed drastically from almost all traffic internal, i.e., wide area networks (WAN) to most of the traffic Internet-based traffic. With this change, our placement of the network security controls also needs to change to avoid traffic congestion and poor performance from existing security controls. The SASE framework (MacDonald, Orans and Skorupa, 2019) addresses the issue of traffic congestion and poor performance to move security controls to the cloud and delivery network security as a service.

The remainder of this paper is organised as follows: Section 2 provides background on cloud services. Section 3 reviews the SASE current state while research gaps are identified in Section 4. The paper concludes with Section 5.

2. Background

This section presents a background on cloud-computing deployment models, cloud-computing service models, everything as a service and the SASE framework and core components are subsequently explained.

2.1 Cloud Overview

According to (Mell and Grance, 2011), "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". In their publication (Mell and Grance, 2011) define cloud computing in terms of four deployment models, i.e. private, community, public and hybrid, and three service models, i.e. Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). These deployment and service models are discussed in more detail in the sections that follow.

2.2 Cloud Deployment Models

A cloud deployment model is described according to where the infrastructure resides and who has control over that infrastructure. A brief description of each follows below:

- A **private cloud** is a cloud infrastructure for restricted use by a single organisation or business.
- A **public cloud** is the most common type of cloud and used when a service provider makes resources, i.e. servers and storage, available to consumers via the Internet.
- A **community cloud** is a joint effort where infrastructure is shared between organisations from a specific community of consumers with the same concerns, e. g. mission, security requirements, policy, and compliance considerations.
- A **hybrid cloud** is a type of cloud infrastructure that results from the combination of two or more distinct cloud infrastructures, i.e., private, community, or public cloud models. A hybrid cloud allows for the flow of data and applications between the different cloud environments.

The next section describes the three cloud services models – Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

2.3 Cloud Service Models

Cloud services are described by (Shilpashree, Patil and Parvathi, 2018) and can be classified in three main categories, IaaS, PaaS and SaaS (see Figure 1). Figure 1 shows that server, storage, networking, security, and physical data centre constitute IaaS. IaaS is a subset of PaaS with PaaS adding additional services such as operating systems, development tools and containers. SaaS adds applications to the list of services and is the topmost layer of IaaS and PaaS.

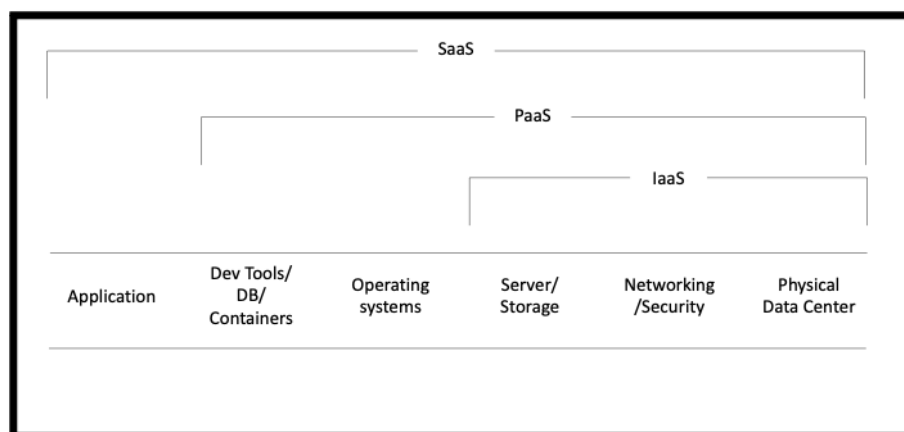


Figure 1: Cloud Services (Stephen Watts, 2019)

- **IaaS** is the first form of cloud computing to provide virtualised resources over the Internet. In an IaaS model, the cloud provider manages the infrastructures such as storage, server and networking resources, while customers must still manage their data, application and operating systems.
- **PaaS** is a cloud computing model that provides an application development environment in the cloud for developers. PaaS is built on top of IaaS and additionally offers services such as middleware, database, containers, etc. for software development. The PaaS customers only manage their applications and data; they are not required to manage the underlying infrastructure or operating system.
- In a **SaaS** cloud computing model, the cloud provider hosts the applications and offers ready-to-use software solutions for businesses. With SaaS, a software vendor may contract a third-party cloud provider to host the application. Larger cloud service providers can also provide all three cloud services.

2.4 Everything as a Service (XaaS)

XaaS is a growing variety of service availability over the Internet via cloud computing, as opposed to on premises. There are countless examples of XaaS, and (Duan, Cao and Sun, 2015) compiled a taxonomy of XaaS. Examples are Storage aaS, Security aaS, Desktop aaS, Database aaS, E-commerce aaS, etc.

This paper focuses on network security aaS and includes firewall as a service (FWaaS), cloud access brokers (CASB), secure web gateways (SWG), zero trust network access (ZTNA) for cloud applications and network as a service (i.e. SD-WAN). A brief description of each is provided below:

- **FWaaS**, also known as cloud firewalls, performs the same functions as a traditional firewall, namely to control traffic sources.
- **CASB** is a cloud-based security policy enforcement point between cloud service consumers and cloud service providers, with the function to block or allow specific applications.
- Cloud **SWG** delivers web security from the cloud to protect web-surfing devices from malware and to enforce corporate compliance.
- **ZTNA** allows only trusted devices access to specific applications and continually authenticates devices and users. A trusted device authentication is not based on IP address, but rather on the identity of the device and the user.
- **SD-WAN** (Software-defined Wide Area Network) is a software-defined approach to the WAN that decouples the network hardware from its control mechanism.

The next section gives an overview of SASE and how the above-mentioned elements for network security as a service are implemented in the SASE framework.

2.5 Secure Access Service Edge (SASE) Overview

One of the most notable frameworks introduced in 2019 was Gartner's cloud network security framework known as SASE. The framework is still in a beginning stage and will probably undergo many changes. (MacDonald, Orans and Skorupa, 2019) gives the following definition of SASE: "The secure access service edge is an emerging offering combining comprehensive WAN capabilities with comprehensive network security functions such as SWG, CASB, FWaaS and ZTNA to support the dynamic secure access needs of digital enterprises." Section 2.5.1 to 2.5.3 discusses why SASE is needed today, SASE's network and security convergence, and a SASE case study.

2.5.1 Why SASE is needed

SASE satisfies the security needs experienced in a highly distributed environment with edge computing, cloud computing, and a workforce that often works from home. Edge computing refers to computing that needs to happen as close as possible to the data to minimise any delays in processing time. Traditionally traffic flows from many different points to a central hub or data centre from where traffic is forwarded to cloud services or elsewhere on the Internet. Experience has shown this design is not efficient, as all traffic is throttled by the central security stack at the data centre. A cloud-native approach – such as using the SASE framework – that does not force traffic through a data centre is then an optimal solution for protecting this new distributed, perimeter-less world.

2.5.2 SASE Network and Security Convergence

With the convergence of network aaS and network security aaS within a SASE deployment. The network services, i.e. SD-WAN, and network security services, i.e. SWG, CASB, ZTNA and FWaaS, are now converged into a single cloud-delivered service model, termed SASE (MacDonald, Orans and Skorupa, 2019). Network and network security teams for support and architectural functions are often different teams in many organisations. For organisations that adopt SASE, there will be a substantial overlap in support and architectural functions between network and security teams. Hence these teams should be collapsed into a single team.

2.5.3 SASE Case Study

Figure 2 shows that a SD-WAN steers all traffic for Internet or cloud services from branch sites towards the SASE security controls. Thus, the SD-WAN also connects the branch offices with the organisation's private data centre. For roaming users not connected to the SD-WAN, the traffic will be steered with an agent on the roaming user device. The SASE cloud comprises its four core components – SWG, CASB, ZTNA and FWaaS – and can identify any malicious traffic from or to branches and roaming users.

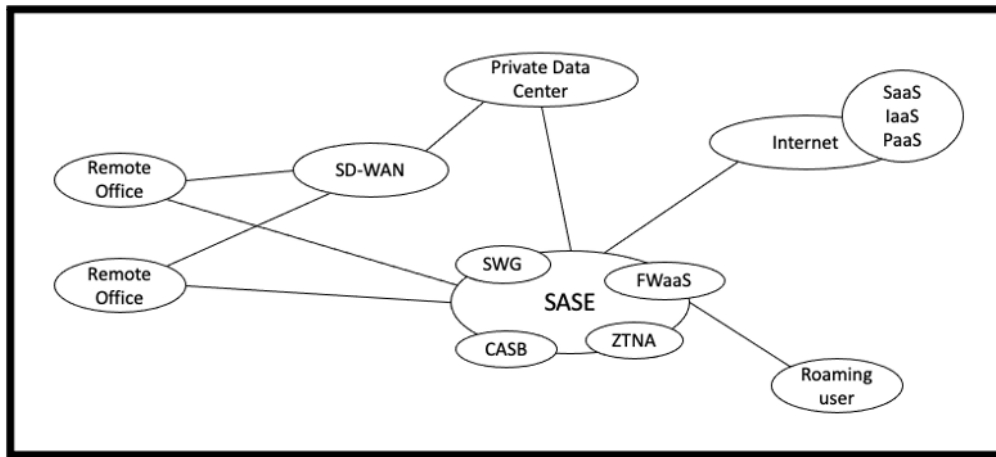


Figure 2: SASE Case study

The next section presents the current state of research with regards to SASE and the five SASE core components.

3. State-of-the-art of SASE

The authors consulted research available on platforms such as the Institute of Electrical and Electronics Engineers (IEEE), Microsoft Academic Research, Google Scholar and the Association for Computing Machinery (ACM) and found only three papers dealing with the topic of SASE. The search was consequently extended to include the SASE core components and numerous papers were found. The SASE core components used to evaluate the state-of-the-art of SASE made use of the SASE framework.

The core components used in the evaluation of SASE state of the art is shown in Table 1. The papers consulted are listed in the left-hand column and an X is entered into the appropriate column to indicate the topic the paper contributes to.

Table 1: Findings

Paper title	1 SASE	2 SD- WAN	3 SWG	4 CASB	5 FWaaS	6 ZTNA	Ref
The future of network security is in the cloud	X						(MacDonald, Orans and Skorupa, 2019)
How SASE is defining the future of network security	X						(Wood, 2020)
MEF White Paper MEF SASE Services Framework July 2020	X	X		X			(MEF Forum, 2020)
Software-Defined Wide Area Network (SD-WAN): Architecture, advances and opportunities		X					(Yang <i>et al.</i> , 2019)
How to make SD-WAN secure		X					(Wood, 2017)
Development and evaluation of a secure web gateway using existing ICAP open-source tools			X				(Pearce and Hunt, 2010)
SD-WAN revolutionises IoT and edge security		X	X				(Pamplin, 2021)
Firewall as a service in SDN OpenFlow network		X			X		(Arins, 2015)
Application-aware firewall mechanism for software-defined networks		X			X		(Nife and Kotulski, 2020)
Towards a standard SDN-based IPsec management framework		X					(Lopez-Millan, Marin-Lopez and

Paper title	1 SASE	2 SD- WAN	3 SWG	4 CASB	5 FWaaS	6 ZTNA	Ref
The future of network security is in the cloud	X						(MacDonald, Orans and Skorupa, 2019)
How SASE is defining the future of network security	X						(Wood, 2020)
MEF White Paper MEF SASE Services Framework July 2020	X	X		X			(MEF Forum, 2020)
Software-Defined Wide Area Network (SD-WAN): Architecture, advances and opportunities		X					(Yang <i>et al.</i> , 2019)
How to make SD-WAN secure		X					(Wood, 2017)
Development and evaluation of a secure web gateway using existing ICAP open-source tools			X				(Pearce and Hunt, 2010)
SD-WAN revolutionises IoT and edge security		X	X				(Pamplin, 2021)
Firewall as a service in SDN OpenFlow network		X			X		(Arins, 2015)
							Pereniguez-Garcia, 2019)
A firewall-adversarial testing approach for software-defined networks					X		(Malkawi <i>et al.</i> , 2021)
Enhancing features of cloud computing using cloud access security brokers to avoid data breaches				X			(Kaur and Gupta, 2019)
Cloud Access Security Broker (CASB): A pattern for secure access to cloud services				X			(Fernandez, Yoshioka and Washizaki, 2015)
CPFirewall: A novel parallel firewall scheme for FWaaS in the cloud environment					X		(Wang <i>et al.</i> , 2015)
Implementation of firewall as a Service for OpenStack Virtualization Systems					X		(Hoang and Bui, 2021)
Relevance of Zero Trust Network Architecture and its rapid adoption amidst Work From Home enforced by COVID-19						X	(D. A. Deshpande, 2021)
Cloud Access Security Brokers (CASBs)				X			(Gartner Inc., 2016)
Enhancing security of cloud platform with Cloud Access Security Broker				X			(Ahmad, Mehruz and Beg, 2021)
Dynamic Cloud Access Security Broker using Artificial Intelligence				X			(Bhattacharya <i>et al.</i> , 2021)
A Study on rapid adoption of Zero Trust Network Architectures by global organizations due to COVID-19 Pandemic						X	(A. Deshpande, 2021)
Beyond Zero Trust: Trust is a vulnerability						X	(Campbell, 2020)
Survivable zero trust for cloud computing environments						X	(Ferretti <i>et al.</i> , 2021)
The top three factors driving zero trust adoption						X	(Embrey, 2020)
Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero trust						X	(Buck <i>et al.</i> , 2021a)

Paper title	1 SASE	2 SD- WAN	3 SWG	4 CASB	5 FWaaS	6 ZTNA	Ref
The future of network security is in the cloud	X						(MacDonald, Orans and Skorupa, 2019)
How SASE is defining the future of network security	X						(Wood, 2020)
MEF White Paper MEF SASE Services Framework July 2020	X	X		X			(MEF Forum, 2020)
Software-Defined Wide Area Network (SD-WAN): Architecture, advances and opportunities		X					(Yang <i>et al.</i> , 2019)
How to make SD-WAN secure		X					(Wood, 2017)
Development and evaluation of a secure web gateway using existing ICAP open-source tools			X				(Pearce and Hunt, 2010)
SD-WAN revolutionises IoT and edge security		X	X				(Pamplin, 2021)
Firewall as a service in SDN OpenFlow network		X			X		(Arins, 2015)
Survey on Zero-Trust Network Security						X	(Yan and Wang, 2020)
A Zero Trust Approach to Network Security						X	(Assunção, 2019)
Latest trend in network security as Zero Trust Security Model						X	(Uttarwar and Kalia, 2019)
Zero Trust Architecture, SP 800-207						X	(NIST, 2020)

The criteria used to evaluate the papers are based on the research relevance to SASE, the number of papers in the mentioned field and gaps in the current SASE research. The next sections briefly discuss the current state of literature of the different topics and its evaluation on each of the components numbered from 1 to 6 in Table 1.

3.1 Secure Access Service Edge (SASE)

Very few contributions were found to be available on the topic of SASE in general, because it is still a very new framework (MacDonald, Orans and Skorupa, 2019). The MEF (originally known as the Metro Ethernet Forum and later re-branded to simply MEF) White Paper (MEF Forum, 2020) defines the MEF SASE Services Framework, which can be used for the implementation of SASE. It provides a way for enterprises and service providers to compare each vendor's approach towards SASE. (Wood, 2020), describes SASE from a SASE provider's point of view and the benefits of SASE.

3.2 Software-defined wide area network (SD-WAN)

A search for SD-WAN returned an extensive list of contributions. A search for "SD-WAN" and "security" on the resulted in fewer returns but is more relevant to this paper and future research on SASE. (Lopez-Millan, Marin-Lopez and Pereniguez-Garcia, 2019) proposes a solution to manage IPsec Security Associations (IPsec SAs) used for encrypting SD-WAN tunnels. The solution by (Lopez-Millan, Marin-Lopez and Pereniguez-Garcia, 2019) can be expanded for SASE framework standardisation (discussed later in Section 4.4).

3.3 Secure Web Gateway (SWG)

The next component is SWG and, again, very little research has been published on this topic. There are also few instances where SWG and SASE are mentioned in the same article. The topic of SWG nevertheless shows many research opportunities within the context of SASE.

3.4 Cloud Access Security Broker (CASB)

The search for "cloud access security broker" returned significant results and the mentioned papers covered a vast diversity of topics. (D Curwin *et al.*, 2021) deal with what a CASB is, while (Obregon, 2017) suggests an approach to use CASB to secure SaaS.

3.5 Firewall as a service (FWaaS)

In the available literature, the authors found only six papers in their search for “FWaaS”. (Hoang and Bui, 2021) proposes the implementation of a firewalling service for cloud systems using OpenStack. (Wei *et al.*, 2016) in turn propose a solution to hide sensitive network policies such as firewall rules from cloud security providers, as they may be leaked and exploited by attackers. The feasibility of implementing the above solution to all SASE components can be investigated.

3.6 Zero trust network access (ZTNA)

Significant research was done in the field of ZTNA over the past two years because the global COVID-19 pandemic and worldwide lockdowns have forced people to work from home. (Wei *et al.*, 2016) confirm the rapid adoption of ZTNA by many organisations due to COVID-19. Legacy hardware VPNs that were designed when a small portion of users were working from home, could not be scaled up for a larger percentage of employees working from home. The ZTNA solution therefore created multiple on-demand micro connections between users’ devices and cloud applications, in order to overcome the scalability issues. (Buck *et al.*, 2021) discuss the flaws of current network security solutions and maintain that the zero-trust model follows the idea that no network – internal or external – can be trusted.

3.7 Summary of current literature

Table 2 gives a summary of the papers identified in SASE state-of-the-art research. The relevance of these papers to SASE (in the authors’ opinion) is indicated in the second column on a scale from 1 to 5 – where 1 means the least significant and 5 means the most significant relevance to SASE. The core components were then sorted from high to low according to their relevance to SASE.

The authors argue that a relevance of 5 is allocated to SASE because these papers focus entirely on Gartner’s SASE framework. The authors allocate a 1 to SD-WAN since SD-WAN, in the context of SASE, is only used to steer traffic to a SASE cloud from branch offices. It, therefore, does not have a significant impact on the organisation of SASE. The role of SWG is to steer traffic to a SASE cloud, and then to monitor and inspect traffic before it is routed to the appropriate cloud service. Since the SWG in a SASE deployment performs several functions, i.e. content filtering, decryption of traffic, data loss prevention for cloud application, and enriches metadata context for investigations, SWG was allocated a 4 out of 5 for the role of SWG in SASE. FWaaS was allocated a 2 since the FWaaS SASE component performs the same role as an on-premises firewall, but with all the benefits of an IaaS offering. ZTNA creates the boundaries to allow only trusted users and devices to access applications. In a SASE deployment, ZTNA becomes a policy enforcement point via a trust broker between users and devices and applications. The trust broker removes applications from public visibility and reduces the attack surface, therefore ZTNA was allocated a 4 out of 5. CASB was rated a 3 because CASB enforces a policy enforcement point between applications and users for security and compliance purposes. Examples of policy enforcements include authentication, authorisation, logging, device profiling and malware detection.

Table 2: Summary of papers

SASE components evaluated	Relevance to SASE 1 - 5	Number of papers
SASE	5	3
ZTNA	4	10
SWG	4	2
CASB	3	6
FWaaS	2	5
SD-WAN	1	7

4. Potential research gaps in SASE

As stated before, the SASE framework converge network and network security aaS. Section 4.1 to 4.4 identifies and discusses gaps in the current research. Section 4.5 summarises the gaps and provides a high-level solution.

4.1 Using SASE as a Managed Detection and Response (MDR) Service for MSSPs

Research is needed to demonstrate how managed security service providers (MSSPs) can use SASE for a managed detection and response (MDR) service. Many large organisations manage their own network and security infrastructure and do not make use of MSSPs; also demonstrated a SASE deployment from an organisational view. In addition to the MDR service, MSSPs also offer Security Operations Centre as a Service (SOCaaS). Due to the convergence of network and security support teams in a SASE environment, the monitoring

of SASE infrastructure makes a convincing case for a combined SOC and NOC. (Miloslavskaya, 2018) describes the advantages of a Network Security Intelligence Centre (NSIC) which is a combination of a Security Intelligence Centre (SIC) and a Network Operations Centre (NOC).

The section that follows discusses the second shortcoming, namely the collection of security events and the transfer of events to a SIEM tool used in a SOC.

4.2 SIEM Integration into SASE via APIs and not Traditional Syslog

Security events are generated by SASE components and maybe a SIEM tool is not needed. Gartner's SASE framework does not address the subject of a security operations centre (SOC) for the monitoring of SASE components. Gartner does however mention application programming interface (API) inspection as a critical capability for a SASE vendor. (MacDonald, Orans and Skorupa, 2019) This implies that some form of monitoring and response capability is needed.

Security events generated by SASE components traditionally send a syslog message to a security information and event management (SIEM) tool via a transport protocol such as syslog or XML. This means that the SASE components forwards the events to a SIEM tool. As an alternative, events can be triggered with an API integration from a SIEM tool to the SASE components and this means that the SIEM tools pulls or fetches the events – as opposed to pushing them. The event logs of Security as a Service (SECaaS) solution are stored with the SECaaS solution provided by the service provider for a short period of time. For long-term event storage, the service provider could store the event data in cloud storage like an Amazon S3 bucket. The benefit of using an API integration as the data source for a SIEM tool is that the number of logs is significantly reduced. Furthermore, the quality of data received by the SIEM tool is improved. Section 4.3 address how to enable digital forensic readiness in the SASE core components.

4.3 Digital Forensic Readiness (DFR) with SASE

Digital forensic readiness (DFR) is still an evolving field but extremely important for saving time and cost in digital forensic investigation. According to (Tan, 2001), digital forensic readiness has two objectives: maximising an environment's ability to collect credible digital evidence and minimising the cost of forensics in an incident response. To recognise the benefits of DFR, more research is needed to construct a framework to ensure all SASE core components are DFR according to ISO/IEC 27043:2015 (Valjarevic, Venter and Petrovic, 2017).

Research done by (Lagrasse *et al.*, 2020) for a framework on DFR in software-defined networks can be applied to be SD-WAN traffic routed towards a SASE cloud. Take into consideration that all SD-WAN tunnel traffic is encrypted with IPsec encryption protocol and as a result all Potential Digital Evidence (PDE) needs to be collected at the edge or hub SD-WAN router. Furthermore, the potential digital evidence collected from the SASE components must be forensically sound and have the potential to withstand legal scrutiny and meet evidence admissibility requirements in a court of law. The DFR framework described by (Lagrasse *et al.*, 2020) uses natural language processing (NLP) to discover and identify cybercrime incidents in the cloud. This technique to use NLP can also be used for a DFR Framework for SASE core components. Section 4.4 discusses the alignment of the SASE framework to existing cloud security standards and frameworks.

4.4 Standardisation of the SASE framework

In our opinion, the Gartner paper on SASE establishes a capability framework that recommends certain functionalities i.e., CASB, FWaaS, ZTNA, SWG and SD-WAN. In our experience multiple vendors build SASE solutions around their existing product capabilities and not according to a standard. Using a standard together with a framework has the following benefits; a common language, a common understanding of requirements, facilitating product integration, enabling regional and international functionality and integration, common technical requirements and cost projections.

To realise the benefits of using standards and tested frameworks, our future research would build on the existing Gartner SASE capability framework by aligning it to existing cloud security standards and frameworks. The next section presents a summary of the gaps discussed and provides a high-level solution to each shortcoming in a table format.

4.5 Summary of research gaps

Table 3 summarises the potential research challenges and suggests a high-level solution to the potential research gaps as identified in Sections 4.1 to 4.4. These solutions should be explored in future research.

Table 3: Summary of research gaps

Potential research gap	Brief description of gap	High-level solution
Using SASE as a MDR service for MSSPs	All SASE components are offered from management service providers (MSPs) and management security service providers (MSSPs). The SASE SD-WAN component has been monitored from a NOC and the other components from a SOC, and the drawback is that they do this disjointly.	Demonstrate how SASE detection and response can be used by MSSPs and how to monitor the SASE core components from a NSIC (i.e. the combination of a SOC and a NOC)
SIEM integration into SASE via APIs and not via traditional syslog	The Gartner SASE framework does not address the subject of a SOC for monitoring SASE.	Use the API capabilities available in SIEM tools to collect security events from the SASE core components (which are also presented through APIs). An API call is sent to an external application to retrieve requested data, i.e. relevant security events, in the SASE components. The benefit of using APIs is that the number of logs is significantly reduced and the quality of data received is improved.
DFR in SASE	Since all SASE components are delivered from the cloud not all of these components are digital forensically ready.	Develop a DFR framework for all SASE core components in order to improve on existing standards.
SASE Framework	In the opinion of the authors, the Gartner SASE framework is already used by multiple vendors and customers; however, it is not aligned to existing standards. This means the SASE framework requires more rigorous testing.	The benefits of using a standard or framework involve the use of a common language; a common understanding of requirements; product integration; regional and international functionality and integration; adherence to common technical requirements; and the facilitation of cost projections. Future research should build on the Gartner SASE capability frameworks and align them to existing cloud security standards and frameworks.

5. Conclusion

The aim of this paper was to establish the present state of development of SASE and to discuss future research opportunities for SASE. An overview was given of cloud services and the different cloud deployment models and cloud service models were discussed. The paper also discussed XaaS (everything as a service) with a focus on network security as a service, i.e. FWaaS, CASB, SWG, ZTNA and SD-WAN, and gave a brief description of each. An overview of the Gartner SASE framework was provided, coupled with a motivation of why SASE is needed today. The convergence of network and network security as a service within in a SASE deployment was discussed and explored in a case study with the SASE framework. This paper then reviewed and evaluated current literature contributions made to SASE and its core components. Our study showed that only scant research literature is currently available on SASE. Therefore, the criteria for reviewed and evaluated literature to be included in the paper were expanded to encompass all the SASE core components.

The paper proceeded to identify four gaps in SASE research. Firstly, the SASE components are monitored disjointly by a NOC and SOC. Secondly, the Gartner SASE framework does not address the matter of a SOC for monitoring of the SASE core components. Thirdly, not all the SASE components are digital forensically ready. Lastly, the SASE framework is not aligned to existing frameworks and standards. Finally, the paper suggested four potential future research endeavours in SASE to address the shortcomings mentioned above. The first topic for future research involved the use of SASE as a managed detection and response (MDR) service for MSSPs. Secondly, SIEM integration into SASE should be done via APIs, and not via traditional logging. Thirdly, digital forensic readiness (DFR) should be incorporated into SASE, and lastly, the SASE framework should be standardised.

References

- Ahmad, S., Mehruz, S. and Beg, J. (2021) 'Enhancing Security of Cloud Platform with Cloud Access Security Broker', in *Lecture Notes in Networks and Systems*. doi: 10.1007/978-981-16-0882-7_27.
- Arins, A. (2015) 'Firewall as a service in SDN OpenFlow network', in *Advances in Information, Electronic and Electrical Engineering, AIEEE 2015 - Proceedings of the 2015 IEEE 3rd Workshop*. doi: 10.1109/AIEEE.2015.7367309.
- Assunção, P. (2019) 'A Zero Trust Approach to Network Security', *Proceedings of the Digital Privacy and Security Conference*.
- Bhattacharya, D. et al. (2021) 'Dynamic Cloud Access Security Broker Using Artificial Intelligence', in *Lecture Notes in Networks and Systems*. doi: 10.1007/978-981-15-7106-0_33.
- Buck, C. et al. (2021a) 'Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust', *Computers & Security*, 110, p. 102436. doi: 10.1016/J.COSE.2021.102436.
- Buck, C. et al. (2021b) 'Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust', *Computers & Security*, 110. doi: 10.1016/j.cose.2021.102436.
- Campbell, M. (2020) 'Beyond Zero Trust: Trust Is a Vulnerability', *Computer*, 53(10). doi: 10.1109/MC.2020.3011081.
- D Curwin et al. (2021) *What is Cloud App Security? | Microsoft Docs, Microsoft*. Available at: <https://docs.microsoft.com/en-us/cloud-app-security/what-is-cloud-app-security>.
- Deshpande, A. (2021) 'A Study on Rapid Adoption of Zero Trust Network Architectures by Global Organizations Due to COVID-19 Pandemic', in *New Visions in Science and Technology Vol. 1*. doi: 10.9734/bpi/nvst/v1/3640f.
- Deshpande, D. A. (2021) 'Relevance of Zero Trust Network Architecture amidst and it's rapid adoption amidst Work From Home enforced by COVID-19', *Psychology and Education Journal*, 58(1). doi: 10.17762/pae.v58i1.2190.
- Duan, Y., Cao, Y. and Sun, X. (2015) 'Various "aaS" of everything as a service', in *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015 - Proceedings*. doi: 10.1109/SNPD.2015.7176215.
- Embrey, B. (2020) 'The top three factors driving zero trust adoption', *Computer Fraud and Security*, 2020(9). doi: 10.1016/S1361-3723(20)30097-X.
- Fernandez, E. B., Yoshioka, N. and Washizaki, H. (2015) 'Cloud Access Security Broker (CASB): A pattern for secure access to cloud services', *4th Asian Conference on Pattern Languages of Programs*.
- Ferretti, L. et al. (2021) 'Survivable zero trust for cloud computing environments', *Computers and Security*, 110. doi: 10.1016/j.cose.2021.102419.
- Gartner Inc. (2016) *Cloud Access Security Brokers (CASBs), Peer Insights*.
- Hoang, X. T. and Bui, N. D. (2021) 'An Implementation of Firewall as a Service for OpenStack Virtualization Systems', in *Lecture Notes in Networks and Systems*. doi: 10.1007/978-981-16-2094-2_12.
- Kaur, S. and Gupta, R. (2019) 'Enhancing Features of Cloud Computing Using Cloud Access Security Brokers to Avoid Data Breaches', *European Journal of Engineering Research and Science*, 4(10). doi: 10.24018/ejers.2019.4.10.1518.
- Lagrasse, M. et al. (2020) 'Digital forensic readiness framework for software-defined networks using a trigger-based collection mechanism', in *Proceedings of the 15th International Conference on Cyber Warfare and Security, ICCWS 2020*. doi: 10.34190/ICCWS.20.045.
- Lopez-Millan, G., Marin-Lopez, R. and Pereniguez-Garcia, F. (2019) 'Towards a standard SDN-based IPsec management framework', *Computer Standards and Interfaces*, 66. doi: 10.1016/j.csi.2019.103357.
- MacDonald, N., Orans, L. and Skorupa, J. (2019) 'The Future of Network Security Is in the Cloud', *Gartner*, (August).
- Malkawi, R. et al. (2021) 'A firewall-adversarial testing approach for software defined networks', *Journal of Theoretical and Applied Information Technology*, 99(1).
- MEF Forum (2020) 'MEF White Paper MEF SASE Services Framework July 2020'.
- Mell, P. and Grance, T. (2011) 'The NIST definition of cloud computing', *Nova Science Publishers, Inc., (2011)*, 171-173, pp. 171-173.
- Miloslavskaya, N. (2018) 'Network Security Intelligence Center as a combination of SIC and NOC', in *Procedia Computer Science*. doi: 10.1016/j.procs.2018.11.084.
- Nife, F. N. and Kotulski, Z. (2020) 'Application-Aware Firewall Mechanism for Software Defined Networks', *Journal of Network and Systems Management*, 28(3). doi: 10.1007/s10922-020-09518-z.
- NIST (2020) 'Zero Trust Architecture, SP 800-207', *National Institute of Standards and Technology Special Publication, SP 800-207*.
- Obregon, L. (2017) *A technical approach at securing SaaS using Cloud Access Security Brokers, SANS*.
- Pamplin, S. (2021) 'SD-WAN revolutionises IoT and edge security', *Network Security*, 2021(8). doi: 10.1016/s1353-4858(21)00090-8.
- Pearce, M. and Hunt, R. (2010) 'Development and evaluation of a secure web gateway using existing ICAP open source tools', in *Proceedings of the 8th Australian Information Security Management Conference*.
- Shilpashree, S., Patil, R. R. and Parvathi, C. (2018) 'Cloud computing an overview', *International Journal of Engineering and Technology(UAE)*, 7(4). doi: 10.14419/ijet.v7i4.10904.
- Stephen Watts, M. R. (2019) *SaaS vs PaaS vs IaaS: What's The Difference & How To Choose – BMC Software | Blogs*. Available at: <https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/> (Accessed: 16 September 2021).
- Tan, J. (2001) 'Forensic Readiness Assessment', *Cambridge, MA: @ Stake*.

- Uttarwar, V. U. and Kalia, A. A. (2019) 'Latest Trend in Network Security as Zero Trust Security Model', *National Journal of Computer and Applied Science*, 2.
- Valjarevic, A., Venter, H. and Petrovic, R. (2017) 'ISO/IEC 27043:2015 - Role and application', in *24th Telecommunications Forum, TELFOR 2016*. doi: 10.1109/TELFOR.2016.7818718.
- Wang, Z. *et al.* (2015) 'CPFirewall: A novel parallel firewall scheme for fwaas in the cloud environment', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-26979-5_9.
- Wei, L. *et al.* (2016) 'A firewall of two clouds: Preserving outsourced firewall policy confidentiality with heterogeneity', in *2016 IEEE Global Communications Conference, GLOBECOM 2016 - Proceedings*. doi: 10.1109/GLOCOM.2016.7841497.
- Wood, M. (2017) 'How to make SD-WAN secure', *Network Security*, 2017(1), pp. 12–14. doi: 10.1016/S1353-4858(17)30006-5.
- Wood, M. (2020) 'How SASE is defining the future of network security', *Network Security*, 2020(12), pp. 6–8. doi: 10.1016/S1353-4858(20)30139-2.
- Yan, X. and Wang, H. (2020) 'Survey on Zero-Trust Network Security', in *Communications in Computer and Information Science*. doi: 10.1007/978-981-15-8083-3_5.
- Yang, Z. *et al.* (2019) 'Software-defined wide area network (SD-WAN): architecture, advances and opportunities', in *Proceedings - International Conference on Computer Communications and Networks, ICCCN*. doi: 10.1109/ICCCN.2019.8847124.