Proceedings of the

18th International Conference on Cyber Warfare and Security

Towson University, Baltimore County Maryland, USA

9-10 March 2023



Richard L. Wilson

and Major Brendan Curran

A conference managed by ACI, UK

Proceedings of the

18th International Conference on Cyber Warfare and Security
ICCWS 2023

A Conference
Hosted By

Towson University, Baltimore County
Maryland, USA

9-10 March 2023

Copyright the authors, 2023. 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.

From 2022 these proceedings are open access and freely available for all to read. The Conference Proceedings for previous years can be purchased from **http://academic-bookshop.com**

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

E-Book ISBN: 978-1-914587-62-7
E-Book ISSN: 2048-9889
Print ISBN: 978-1-914587-61-0
Book Version ISSN: 2048-9870

Published by Academic Conferences International Limited

Reading, UK

+44118 324 6938

www.academic-conferences.org
info@academic-conferences.org

ICCWS 2023 Contents Page

Preface **v**

Committee **vi**

Biographies **vii**

Academic Papers

Secure Cloud Migration Strategy (SCMS): A Safe Journey to the Cloud
Dalal Alharthi 1-6

Cyber-Physical Attack Using High Power RF in Havana, Cuba
Allyffazzkkamn Argudo, Ghislaine Nasibu 7-18

Cybersecurity in Digital Transformation applications: Analysis of Past Research and Future Directions
Zakariya Belkhamza 19-24

An Automated Post-Exploitation Model for Cyber Red Teaming
Ryan Benito, Alan Shaffer, Gurminder Singh 25-34

Review of End-to-End Encryption for Social Media
Vijay Bhuse 35-37

An Analysis of Crypto Scams during the Covid-19 Pandemic: 2020-2022
Johannes George Botha, Danielle Botha, Louise Leenen 36-48

Case Study: Conducting a Risk Assessment for an Electrical Utility
Edwin Covert 49-56

On the Use and Strategic Implications of Cyber Ranges in Military Contexts: A Dual Typology
Mischa Hansel, Andrew Dwyer, Kathrin Moog, Jantje Silomon 57-66

Evaluation of Quantum Key Distribution for Secure Satellite-integrated IoT Networks
Andrew Edwards, Yee Wei Law, Ronald Mulinde, Jill Slay 67-76

Commentary on Healthcare and Disruptive Innovation
Hilary Finch, Abasi-Amefon Affia, Woosub Jung, Lucas Potter, Xavier-Lewis Palmer 77-84

Securing Commercial Satellites for Military Operations: A Cybersecurity Supply Chain Framework
Courtney Fleming, Mark Reith, Wayne Henry 85-92

Predictors of Human Efficiency in Radar Detection Tasks
Elizabeth Fox, Arielle Stephenson, Christopher Stevens, Gregory Bowers 93-102

Nuclear Cyber Attacks: A Study of Sabotage and Regulation of Critical Infrastructure
Virginia Greiman 103-110

Social-Engineering, Bio-economies, and Nation-State Ontological Security: A Commentary
Brandon Griffin, Keitavius Alexander, Xavier-Lewis Palmer, Lucas Potter 111-118

Search and CompAre Reverse (SCAR): A Bioinformatics–Inspired Methodology for Detecting File Remnants in Digital Forensics
George Grispos, William Mahoney, Sayonnha Mandal 119-127

Development and Analysis of a Reconnaissance-Technique Knowledge Graph
Thomas Heverin, Elsa Deitz, Eve Cohen, Jordana Wilkes 128-136

Zero Trust is Not Enough: Mitigating Data Repository Breaches
JS Hurley 137-144

Managing Large-Scale Heterogeneous Deployments for Cybersecurity
JS Hurley 145-151

Digital Geopolitics: A Review of the Current State
Gazmend Huskaj 152-161

Managing Variable Cyber Environments with Organizational Foresight and Resilience Thinking
Eveliina Hytönen, Jyri Rajamäki, Harri Ruoslahti 162-170

Fingerprinting Network Sessions for the Discovery of Cyber Threats
Christiaan Klopper, Jan 171-180

S-400s, Disinformation, and Anti-American Sentiment in Turkey
Russell Korb, Saltuk Karahan, Gowri Prathap, Ekrem Kaya, Luke Palmieri, Hamdi Kavak 181-188

Detecting and tracking hypersonic glide vehicles:A cybersecurity-engineering analysis of academic literature
Yee Wei Law, John Joshua Gliponeo, Dilpreet Singh, John McGuire, Jiajun Liang, Sook-Ying Ho, and Jill Slay 189-198

Cyber security training in Finnish basic and general upper secondary education
Martti Lehto, Pekka Neittaanmäki 199-208

Russian Influence Operations during the Invasion of Ukraine
Joseph Littell, Nicolas Starck 209-217

Modelling the Influential Factors Embedded in the Proportionality Assessment in Military Operations
Clara Maathuis, Sabarathinam Chockalingam 218-226

Social Media Manipulation Awareness through Deep Learning based Disinformation Generation
Clara Maathuis, Iddo Kerkhof 227-236

Social Media Manipulation Deep Learning based Disinformation Detection
Clara Maathuis, Rik Godschalk 237-245

Improvements on Hiding x86-64 Instructions by Interleaving
William Mahoney, Todd McDonald, George Grispos, Sayonnha Mandal 246-255

Developing Privacy Incident Responses to Combat Information Warfare
Sean McElroy, Lisa McKee 256-263

Use of Intrusion Detection Systems in Vehicular Controller Area Networks to Preclude Remote Attacks
Anthony Monge, Todd Andel 264-272

Digital Insanity: Exploring the Flexibility of NIST Digital Identity Assurance Levels
Kenneth Myers 273-278

Risk likelihood of planned and unplanned cyber-attacks in small business sectors: A cybersecurity concern
Tabisa Ncubukezi 279-290

Ret-gadgets in RISC-V-based Binaries Resulting in Traps for Hijackers
Toyosi Oyinloye, Lee Speakman, Thaddeus Eze 291-299

Basic Elements of Cyber Security for a Smart Terminal Process
Jouni Pöyhönen, Jussi Simola, Martti Lehto 300-308

Anti-American Stance in Turkey: A Twitter Case Study
Gowri Prathap, Hamdi Kavak, Ekrem Kaya, Luke Palmieri, Saltuk Karahan, Alex Korb 309-317

Gaps in Asset Management Systems to Integrate Railway Companies' Resilience
Jyri Rajamäki, Jari Savolainen, Rauno Pirinen, Villamor 318-326

LoRaWAN & The Helium Blockchain: A Study on Military IoT Deployment
Michael A. Reyneke, Barry E. Mullins , Mark G. Reith 327-337

Blockchain Forensics: A Modern Approach to Investigating Cybercrime in the Age of Decentralisation
Saminu Salisu, Velitchko Filipov 338-347

Biocybersecurity and Deterrence: Hypothetical Rwandan Considerations
Issah Samori, Gbadebo Odularu, Lucas Potter, Xavier-Lewis Palmer 348-354

Lesson Plan: An Interdisciplinary Approach to Teaching Cyber Warfare Concepts
Donna Schaeffer, Patrick Olson 355-359

The Impact of Edge Computing on the Industrial Internet of Things
Nkata Sekonya, Siphesihle Sithungu 360-368

Cyber Threat Analysis in Smart Terminal Systems
Jussi Simola, Jouni Pöyhönen, Lehto Martti 369-378

Towards a Scientific Definition of Cyber Resilience
Sidney Smith 379-386

Implications of Cyberbiosecurity in Advanced Agriculture
Simone Stephen, Keitavius Alexander, Lucas Potter, Xavier-Lewis Palmer 387-393

Organizational Cybersecurity Post The Pandemic: An Exploration of Remote Working Risks and Mitigation Strategies
Stephen Treacy, Anoop Sabu, Thomas Bond, Joseph O'Sullivan, Jack Sullivan, Peter Sylvester 394-401

A Quantitative Risk Assessment Framework for the Cybersecurity of Networked Medical Devices
Maureen Van Devender, Jeffrey Todd McDonald 402-411

Towards the Usefulness of Learning Factories in the Cybersecurity Domain
Namosha Veerasamy, Thuli Mkhwanazi, Zubeida Dawood 412-419

Naïve Bayes Supervised Learning based Physical Layer Authentication: Anti-Spoofing techniques for Industrial Radio Systems
Andreas Weinand, Christoph Lipps, Michael Karrenbauer,Hans Dieter Schotten 420-430

How the Russian Influence Operation on Twitter Weaponized Military Narratives
Dana Weinberg, Jessica Dawson, April Edwards 431-439

Nuclear Weapons, Cyber Warfare, and Cyber Security: Ethical and Anticipated Ethical Issues
Richard Wilson, Alexia Fitz 440-448

Robots Security Assessment and Analysis Using Open-Source Tools
Benjamin Yankson, Tyler Loucks, Andrea Sampson, Chelsea Lojano 449-456

Social Robots Privacy Enhancement Using Colored Petri Net (CPN) for Behavior Modeling: A Case Study of Asus Zenbo Robot
Benjamin Yankson, Farkhund Iqbal, Fadya AlMaeeni 457-464

PhD Papers

A Unified Forensics Analysis Approach to Digital Investigation
Ali Alshumrani, Nathan Clarke, Bogdan Ghita 466-475

Offensive Cyberspace Operations for Cyber Security
Gazmend Huskaj 476-479

Categorizing Cyber Activity Through an Information-psychological and Information-technological Perspective, Case Ukraine.
Harry Kantola 480-488

Using Military Cyber Operations as a Deterrent
Maria Keinonen 489-496

Identifying Commonalities of Cyberattacks Against the Maritime Transportation System
Rebecca Rohan 497-503

Evaluating a Non-platform-specific OCR/NLP system to detect Online Grooming
Jake Street, Funminiyi Olajide 504-511

Masters Papers

Demonstrating Redundancy Advantages of a Three-Channel Communication Protocol
Scott Culbreth, Scott Graham 513-522

UAV Payload Identification with Acoustic Emissions and Cell Phones
Hunter Doster, Barry Mullins 523-533

Just Warfare: Is a Nuclear Attack an appropriate Response to a Cyber Attack?
Alexia Fitz, Richard Wilson 534-541

A Review and Testing of Fault Tolerance Levels of Anti-Poaching Cybersecurity System
Isabelle Heyl, Julia Stone, Takudzwa Vincent Banda, Vian Smit, Dewald Blaauw 542-549

DACA: Automated Attack Scenarios and Dataset Generation
Frank Korving, Risto Vaarandi 550-559

Towards a Critical Review of Cybersecurity Risks in Anti-Poaching Systems
Christelle Steyn, Dewald Neville Blaauw 560-568

Work in Progress Papers

Locality-based electromagnetic leakage assessment using CNN
Ian Heffron, James Dean 570-576

Using Deep Reinforcement Learning for Assessing the Consequences of Cyber Mitigation Techniques on Industrial Control Systems
Terry Merz, Romarie Morales Rosado 577-580

ICCWS Preface

These proceedings represent the work of contributors to the 18th International Conference on Cyber Warfare and Security (ICCWS 2023), hosted Towson University, Baltimore County, Maryland, USA on 9-10 March 2023. The Conference co-Chairs are Richard Wilson, Towson University, Baltimore County, Maryland, USA, and Major Brendan Curran, from National Security Agency's Space Cybersecurity Office in Fort Meade, Maryland.

ICCWS is a well-established event on the academic research calendar and now in its 18th 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 Justin Fanelli, Technical Director, Dept of Navy PEO Digital, on the topic of *Can Zero Trust Restore Our Ailing Trust?* The second day of the conference will open with an address by Dr. George R. Lucas, Stockdale Center at the U.S. Naval Academy. His talk is titled *Military Ethics & Algorithmic Warfare: from the Battlefield to the Cyber Domain*

With an initial submission of 136 abstracts, after the double blind, peer review process there are 53 Academic research papers, 6 PhD research papers, 12 Masters Research papers and 1 work-in-progress paper published in these Conference Proceedings. These papers represent research from Brazil, Austria, China, Estonia, Finland, Germany, Israel, Italy, Kenya, Nigeria, Pakistan, Russia, Saudi Arabia, Slovakia, South Africa, Sweden, Switzerland, Taiwan, UAE, Uganda and the United Kingdom.

We hope you enjoy the conference.

Richard L. Wilson
Towson University, Baltimore County
Maryland, USA
March 2023

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 Andel, 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 Mustafa Canan, Naval Postgraduate School, USA; Dr Kevin Matthe Caramancion, Merceyhurst University, Erie, PA, USA; Dr Jim Chen, U.S. National Defense University, USA; Dr Sabarathinam Chockalingam, Institute for Energy Technology, Norway; 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 Jonathan Fuller, United States Military Academy, 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 George Grispos, University of Nebraska at Omaha, USA; 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 Ahmet Koltuksuz, Yasar University, Turkey; Mr Peter Kunz, Hexagon Geosystems, Germany; Dr Gregory Laidlaw, University of Detroit Mercy, USA; 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; Christoph Lipps, German Research Center for Artificial Intelligence,Germany; Volodymyr Lysenko, University of Washington, USA; Dr Bill Mahoney, University of Nebraska, Omaha, USA; ASSI Haribabu Maruturi, qiscet, india; Dr Paul Maxwell, Army Cyber Institute, USA; Dr Todd McDonald, Air Force Institute of Technology, USA; Dr Terry Merz, State University of New York College of Emergency Preparedness, Homeland Security and Cybersecurity, United States; Dr Robert Mills, Air Force Institute of Technology, USA; Assc Pardis Moslemzadeh Tehrani, Faculty of Law. University of Malaya, Malaysia; Dr Barry Mullins, Air Force Institute of Technology, USA; Prof Antonio Muñoz, University of Málaga, Spain; Dr Asoke Nath, St. Xavier's College(Autonomous), India; Dr Funminiyi Olajide, Nottingham Trent University, UK; Prof Abdelnaser Omran, School of Economics, Finance and Banking, Universiti Utara Malaysia, Malaysia; Rain Ottis , Tallinn University of Technology, Estonia; Dr Jaydeep Patil, D.Y.Patil Agriculture and Technical University Talsande, India; 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, Petoria, 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-Nata, 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 Yilun Shang, Singapore University of Technology and Design, Singapore; Dr D. Cragin Shelton, Capitol Technology University, USA; Dr Dan Shoemaker, Centre for Assurance Studies, USA; Mr Avinash Singh, University of Pretoria, South Africa; Prof Jill Slay, University of South Australia, Australia; 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; Dr Le Thi-Thu-Huong, Pusan National University, Korea; 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; 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



Richard L. Wilson is a Professor in Philosophy at Towson University in Towson, MD. Professor Wilson Teaches Applied Ethics in the Philosophy and Computer and Information Sciences departments at Towson while also serving as Senior Research Scholar in the Hoffberger Center for Professional Ethics at the University of Baltimore. Professor Wilson's interest's are directed towards applying phenomenology to issues with emerging and innovative technologies.



Major Brendan Curran currently serves as a strategist for the National Security Agency's Space Cybersecurity Office in Fort Meade, Maryland. Major Curran is an Army Space Operations Officer with over twenty years of experience as both an active duty U.S. Army Commissioned and Non-Commissioned Officer in the fields of engineering, intelligence, communications, and space operations. Over the course of his fourteen years in the space community Brendan has integrated space-based capabilities and emerging technologies in support of military activities throughout Europe, Southwest Asia, and Africa. Major Curran established the Space Operations & Integrated Joint Special Technical Operations programs at 10th Special Forces Group (Airborne), Mission Command Training Program, and the Army Space Officer position at U.S. Space Force Space Operations Command (then U.S. Air Force Space Command) where he co-chaired the initial planning team that led to the creation of the National Space Defense Center. His professional passions are operational design, unconventional warfare, and strategy. Brendan is Defense Advanced Research Projects Agency (DARPA) Service Chief Fellow, Horatio Alger Association Distinguished Military Scholar, recipient of the BG Sharp Distinguished Leadership Award, and alumni of the U.S. Army training with industry program at the Johns Hopkins University Applied Physics Laboratory. He attended Towson University, MD on an Army Green to Gold scholarship where he earned a BS in Philosophy. Major Curran later attended Norwich University where he earned a MA in Diplomacy with a concentration in international conflict resolution. Major Curran is married to Dr. Carrie Arnold. They have a seven-year-old daughter. When he's not gazing up at the stars, Brendan can be found immersed in gardening and herpetology.

Keynote Speakers



Mr. Justin Fanelli is the Technical Director of Program Executive Office Digital & Enterprise Services (PEO Digital), chartered to provide strategic technical leadership for the rapid transformation of enterprise IT infrastructure within the Department of Navy. His portfolios include Cloud and Platforms, Cyber and Operations, Digital Workplace/Productivity, Infrastructure, End User Services, and several others. Mr. Fanelli also serves as Chief Architect of Defense Medical Intelligence Data and Advisor to The Advanced Research Projects Agency for Health (ARPA-H), the moonshot startup within US Department of Health and Human Services.



Dr. George R. Lucas is a specialist in applied ethics (military, professions, technology) and history of modern philosophy/philosophy of science. He is Distinguished Chair in Ethics Emeritus at, Stockdale Center for Ethical Leadership, U.S. Naval Academy (Annapolis MD), Emeritus status conferred by Secretary of the Navy in April 2014 for life. Lucas is widely published, most recently with the book, Law, Ethics, and Emerging Military Technologies: Confronting Disruptive Innovations (Oxford: Routledge, 2022). He is also the General Editor of a Collected Critical Edition of the Works of Alfred North Whitehead (Edinburgh University Press). Lucas' current areas of research interest are Cyber Ware Ethics and AI and Cyber Warfare.

Mini Track Chairs



Dr. Jawad Hussain Awan is Assistant Professor, Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology, Gharo Campus, Sindh, Pakistan. He 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 & ECCWS 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. Zakariya Belkhamza is currently a member of faculty at the computer information Systems department, Ahmed Bin Mohammed Military College, Doha, Qatar. His area of expertise is information systems, strategic information systems, system design and architecture. Dr. Zakariya Belkhamza holds a Ph.D. in Information systems. He has published several papers in International journals and has authored and edited several books in the area of information systems.



Dr. Jason C. Brown is a research scientist and assistant professor at the Army Cyber Institute at West Point. He teaches risk management and organizational security, and as a futurist, he studies emerging threats, technological and social trends, and responses to those threats. A currently serving officer with the U.S. Army, Dr. Brown has worked within the intelligence, information operations, and cyber career fields. He has authored technical reports on the future of extremism, information warfare, cyber enabled financial crimes, and Chinese soft power.



Dr. Jessica Dawson is the Information Warfare Division Chief at the Army Cyber Institute at West Point. She holds a Ph.D. from Duke Sociology and teaches social inequality and social theory. Her research focuses on the digital disruption of social processes. She serves as a special advisor to Army senior leaders on social media threats and information warfare.



Christoph Lipps, graduated in Electrical and Computer Engineering at the University of Kaiserslautern, where he meanwhile lectures as well. He is a Senior Researcher and Ph.D. candidate at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, Germany, heading the Cyber Resilience & Security Team of the Intelligent Networks Research Department. He is not only a member of the committee of the ICCWS and ECCWS, but also scientific committee member of various international conferences, and reviewer of several journals. His research focuses on Physical Layer Security (PhySec), Physically Unclonable Functions (PUFs), Artificial Intelligence (AI), entity authentication, Security in the Sixth Generation (6G) Wireless Systems and all aspects of network and cyber security.



Dr. Char Sample is a researcher with ICF. Dr. Sample also teaches at Boise State University. She has given guest lecturers at Bournemouth University, Rensselaer Polytechnic University and Royal Holloway University. Dr. Sample has over 20 years' experience in the information security industry. Dr. Sample's research focuses on deception, and the role of cultural values in cybersecurity events **Update**. More recently she has begun researching the relationship between human cognition and machines. Presently Dr. Sample is continuing research on modeling cyber behaviors by culture, other areas of research are data resilience, cyber-physical systems and industrial control systems.



Dr. Pardis Moslemzadeh Tehrani, is a Visiting Associate Professor at the Faculty of Law, the University of Malaya. She was a Senior Lecturer at the same university from 2015 to 2022. Her research interests span the area of Information Technology Law, International Humanitarian Law and legal research methodology. Pardis has made scholarly contributions in peer-reviewed journals and presented papers in several national and international conferences and served on many conferences and workshop program committees.



Dr Stilianos Vidalis, Head of the Networks, Security and Systems subject group at the Department of Computer Science of the University of Hertfordshire. Leading the development of an innovative technology that automates threat assessments. Experienced in all aspects of Information Operations (IO) with excellent understanding of the Information Security (IS) lifecycle. Acted as an instructor for the British Armed Forces Intelligence Personnel on penetration testing and digital forensics for a number of years. Acted as a cyber-expert for the Welsh Government and for the Wales University Officers' Training Corps (WUOTC).

Biographies of Contributing Authors

Azza M. AlGhamdi is a Cybersecurity doctoral candidate at Marymount University. Her dissertation entitled "Establishing a Cybersecurity School to Close the Workforce Gap" explores how to make cybersecurity a more visible discipline and domain of study in universities and how to attract more college freshmen to pursue cybersecurity as a career to increase student enrollment.

Allyffazzkkamn R. Argudo, 19 years of active duty service (U.S. Army), working on satellite communications / cybersecurity. I'm working on my Doctor of Science, Cybersecurity. Research focuses on data transmission security thru multi-spectrum orchestration. I hold a masters in information assurance (Thomas Edison State University, 2020) and a bachelor's in electronics engineering (2017).

Ali Alshumrani is pursuing his Ph.D. in Digital Forensics at the Centre for Cyber Security, Communications and Network Research (CSCAN) at the University of Plymouth. His research is focused on developing a unified digital forensics tool to advance data analytics automation. His research interests reside in digital forensics, cyber security, data ontology and data graph.

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, Malware Analysis and Secure Coding.

Micki Boland is a global cybersecurity warrior and evangelist with Check Point Software Technologies Office of the CTO. A practitioner with 20 years in IT, cybersecurity, emerging technology innovation, Micki holds ISC2 CISSP, Master of Science in Technology Commercialization University of Texas at Austin and MBA Global Security Concentration from East Carolina University.

Johnny Botha enrolled for a PhD in Computer Science at the University of Western Cape. Topic: *"A process for Blockchain Crime Investigations Based on Open-source Intelligence".* Project Manager & Researcher at the CSIR. Masters degree in Information Technology, at University of South Africa(UNISA). NDip and BTech degree in Computer Systems Engineering at the Tswane University of Technology(TUT).

Edwin Covert is a cybersecurity director for a large media company. Previously, he was the Global Head of Security Architecture for a global bank. He has a Bachelor of Science in cybersecurity from the University of Maryland University College. He is currently a Master's candidate at Colorado State University studying the management of information technology.

Scott Culbreth is a Second Lieutenant in the United States Air Force. He graduated from the United States Air Force Academy in 2021 with a B.S. in Computer Science and the Air Force Institute of Technology in 2022 with a M.A.S. in Cyber Systems. Research Interests include Secure Networking and Defense focused technologies.

Hunter Doster is an Air Force Officer currently working at the Air Force Research Lab. He graduated from the United States Air Force Academy in 2020 with a B.S. in computer engineering and from the Air Force Institute of Technology in 2022 with a M.S. in Computer Engineering. His main research areas are digital signal processing for acoustic emissions and digital RF systems.

Andrew Dwyer is a Lecturer (Assistant Professor) in Information Security at Royal Holloway, University of London. His research spans computational decision-making, critical approaches to cybersecurity, as well as exploring the role of cyber operations in contemporary conflict. He completed his DPhil in Cyber Security at the University of Oxford in 2019.

Andrew Edwards is a student at the University of South Australia, currently undertaking a Master of Telecommunication Engineering. The focus of research for his minor thesis is in quantum cryptography as applied to satellite-integrated communication networks.

Sulemana Bankuoru Egala is a Doctoral researcher in Management Science and Engineering majoring in data mining and information management at the School of Management and Economics, University of Electronic Science and Technology of China (UESTC) and an Assistant lecturer at the Department of Informatics, SD Dombo University of Business and Integrated Development Studies, Wa, Ghana.

Hilary Finch is a masters student at Old Dominion University at the School of Cybersecurity. As a former educator of almost 10 years, she brings a unique perspective to the importance of research in cybersecurity. She currently serves as the Educational Liaison for CySecSol, LLC, an organization that brings underrepresented communities into the field of cybersecurity.

Alexia Fitz is an undergraduate student at Towson University dual majoring in International Studies and Political Science with a minor in Human Rights and History. She has found a passion for researching laws of war and WMDs. She plans on continuing her academic career with a master's in security studies, followed by law school.

Courtney Fleming attends the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, OH studying for her Master's degree in Cyber Operations. She graduated from Mississippi State University with a Bachelor's degree in Mathematics in 2018. Her research interests include space cybersecurity, cyber attack tactics, and software vulnerability analysis.

Elizabeth Fox is a research psychologist at the Air Force Research Laboratory, Wright-Patterson Air Force base, Ohio. She received her PhD in Human Factors and Industrial-Organisational Psychology from Wright State University in 2019. She develops and applies models of performance and neural activity to assess cognitive processes and inform system design and automation.

Virginia A. Greiman, Assistant Professor (Global Cyber Law and Governance and Megaprojects / Infrastructure Planning, Boston University). Served as a U.S. Department of State delegation diplomatic head in Eastern/ Central Europe, Asia and Africa and held high level appointments with the U.S. Department of Justice. Researches and publishes: cyber resilience, cyber law/regulation, nuclear cyber security, and global infrastructure megaprojects.

Brandon Griffin is a current cybersecurity student with CySecSol. He has received a bachelor's degree in kinesiology from Old Dominon University in Norfolk, Virginia. Brandon is interested in cyberbiosecurity, informational warfare, social media and the relationship they have on national security.

George Grispos - Assistant Professor of Cybersecurity (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: digital forensics, cybersecurity, residual data, and incident response. Has conducted research with Fortune 500 Organizations (financial services and manufacturing sectors), and law enforcement agencies.

Mischa Hansel leads IFSH's "International Cybersecurity" (ICS) research focus. His research interests encompass cybersecurity strategies and military postures as well as multilateral efforts to regulate offensive cyber operations. He completed his PhD with a thesis on conflict and cooperation in the field of international cybersecurity at the University of Cologne.

Ian Heffron is a Captain in the United States Air Force and currently attends the Air Force Institute of Technology. He has a background in computer science and serves as a Cyberspace Operations Officer. He has served overseas in South Korea and in Germany.

Thomas Heverin is the lead of Girls Learn Cyber, Inc., a training company for K-12 girls in cybersecurity. He holds a Ph.D. focused on cybersecurity, the CISSP, and years of cybersecurity experience. He led research and development projects for the U.S. Navy after serving as a Surface Warfare Officer.

Isabelle Heyl, is currently a student at the University of Stellenbosch. I grew up in Johannesburg and moved to Stellenbosch in 2019. I am 21 years old and passionate about my studies and my hobbies such as triathlon and guitar.

J.S. Hurley joined ODNI in December 2020 focusing on the use of advanced data science tools and techniques and cyber intelligence. Hurley previously served as Professor at NDU in CIC focused on data analytics and cyberspace strategies and formerly as Senior Manager, Distributed Computing, The Boeing Company, in the Network Systems Division.

Gazmend Huskaj: Head of Cyber Security (GCSP). Previously, doctoral student in Cyberspace Operations (Swedish Defence University). Was Director of Intelligence on Cyber-related issues in the Swedish Armed Forces. A veteran five years plus of duty in conflict and post-conflict areas including two Balkan tours and one in Central Asia. Current work: capacity-building, dialogue, and policy analysis.

Harry Kantola (LtCol) is currently appointed as Research Manager at Finnish Defence Research Agency. He holds a General staff officers' degree (Licentiate) from the Finnish National Defence University, a Master's degree in War studies from the Swedish National Defence University and a Post Graduate certificate in Cyber Defence and Information Assurance from Cranfield University, UK.

Hamdi Kavak is an Assistant Professor in the Computational and Data Sciences Department and co-director of the Center for Social Complexity at George Mason University. His research interests lie at the intersection of data science and modeling & simulation. His email and website addresses are hkavak@gmu.edu and http://www.hamdikavak.com.

Maria Keinonen is a senior research officer and a PhD student for Military Sciences in the National Defence University in Finland. In her doctoral research, she focuses on the study of cyber deterrence from the perspective of a small state.

Christiaan Klopper is a master's student in IT focusing on big data science at the University of Pretoria, SA. He received his BEng in electronic engineering from the University of Pretoria in 2010. His main research areas are data science, big data analytics and developing a self-learning cyber defence system that can discover undetectable threats.

Alex Korb is a PhD student and graduate research assistant at Old Dominion University, Virginia, USA. He received his Master's degree in International Studies in 2019. His primary research interests are cyber- and information warfare, globalization and internet culture, and disinformation.

Frank Korving obtained his Master's degree at Tallinn University of Technology, Estonia in 2022. He has been a security practitioner for a large Nordic financial institution for the last five years.

Gentry Lane is the CEO & Founder of ANOVA Intelligence, a cyber national security software company. She is a senior software executive, statistician and cyberpower military strategist. She holds a DoD-appointed position to NATO STO, is a Senior Fellow at the Potomac Institute & Fellow at the National Security Institute.

Yee Wei Law is a Senior Lecturer at UniSA STEM, University of South Australia. His research is currently focusing on space system security, adversarial machine learning and machine diagnostics.

Martti Lehto, PhD (Military Sciences), Col (GS) (ret.): Cyber security (University of Jyväskylä). Over 30 years of experience as developer and leader of C4ISR Systems in Finnish Defence Forces. He has over 200 publications, research reports and articles on the areas of C4ISR systems, cyber security and defence, information warfare, air power and defence policy.

Joe Littell (joseph.littell@westpoint.edu): US Army psychological operations officer and research scientist at the Army Cyber Institute at the United States Military Academy. His research focus on information warfare and disinformation. He holds degrees a bachelor's degree in Computer Science from the University of South Florida and a master's degree in Data Science from Duke University.

Rich Losier is a high-energy, leader driving collaborative security innovation. He has a proven record of rapid prototyping for strong business results. He is a thought leader for DevOps in preparation for IP licensing or M&A activity. Through personal values of integrity, respect, courage and commitment, Rich is a perpetual learner, team-builder, and positive organizational force.

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.

Dwayne McDaniel has been working as a Developer Relations professional since 2015 and has been involved in tech communities since 2005. He loves sharing his knowledge, and he has done so, by giving talks at over a hundred events worldwide. Dwayne currently lives in Chicago. Outside of tech, he loves karaoke, live music, and performing improv.

Sean McElroy has built financial services products throughout his 20-year career. He serves as the CSO of Lumin Digital, a digital banking fintech. Previously, he co-founded Alkami Technology. Sean is a Ph.D. student in Dakota State University's Cyber Defense program and earned a Masters of Science in Information Security Engineering from the SANS Technology Institute.

Lisa McKee, a highly regarded security and privacy expert, has 20 years of industry experience, and a regular featured speaker at conferences and events globally including RSAC. Her Ph.D. is in Cyber Defense and dissertation in Privacy from Dakota State University where she shares her passion for privacy and security as an associate professor.

Anthony Monge is a retired US Naval officer and currently a PhD candidate at the University of South Alabama in Mobile. His current research is in Controller Area Network (CAN) Bus security for vehicles and similar Cyber-Physical systems. He currently resides in the Dallas-Fort Worth area with his wife and children.

Kenneth Myers is a doctoral student in the cybersecurity program at Marymount University, Arlington, VA, USA. His main research areas are identity and access management, cybersecurity workforce planning, and digital identity.

Ghislaine Nasibu Mwana, I am currently working toward a Doctor of Science in Cybersecurity. My thesis dissertation topic is "Cybersecurity consideration when adopting blockchain technology in the accounting field". I have a Master of Accountancy degree from the George Washington University. My researches are oriented toward how technology can impact the financial sector and processes.

Tabisa Ncubukezi is an academic at the Cape Peninsula University of Technology. She has done her PhD in the same institution. Her primary research areas are communication networks, cloud computing, cyber security, and e-Learning.

Victor Kwarteng Owusu is a Ph.D. candidate at the Faculty of Management and Economics, Tomas Bata University in Zlin, Czech Republic. He has two publications to his credit. He has also attended conferences across Europe and the United Kingdom including British Academy

of management (BAM), hosted by Manchester business school. His main research area is organizational performance.

Toyosi Oyinloye is a lecturer at the University of Chester, UK. Her main research area is in software exploitation and software protection methods which she has been studying for the past three years in her doctorate. She has authored other related papers as outcomes of her study.

Ada Peter is a Professor, Digital Media, and International Security. She explores the processes of cyber weaponization, terrorism, and conflict to help governments make informed decisions about warfare in the fifth domain.

Gowri Prathap holds a bachelor of science degree in Computational and Data Sciences from George Mason University (GMU). She is currently a TECDP analyst working at Cigna. Before joining Cigna, she worked as a research and teaching assistant at GMU. Her research interests are Artificial Intelligence and Machine Learning.

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 the University of Jyväskylä, Finland. He holds D.Sc. degrees in electrical and communications engineering from the Helsinki University of Technology, and a PhD in mathematical information technology from the University of Jyväskylä.

Michael A. Reyneke received his B.S. in Computer Engineering from Utah Valley University in 2017. He is currently a Cyber Operations masters' degree student at the Air Force Institute of Technology. His current research interests include the Internet of Things, wireless communications, critical infrastructure protection, cyber-physical systems security/sensing, national security, and reverse engineering.

Rebecca Rohan is a resident in the Doctorate of Science in Cybersecurity program at Marymount University, Arlington, Virginia, U.S. She received a MS in cybersecurity with concentrations in forensics and intelligence from Utica University, Utica, NY, in 2012. Her main research areas are maritime cybersecurity, cybersecurity education, cybersecurity intelligence, and diversity in cybersecurity.

Saminu Salisu is the CTO of Bilic.io, and a PhD student at Vienna University of Technology, Austria. His research interests are Information Visualisation, Visual Analytics, Uncertainty Visualisation and Spatio-Temporal Visualisation. He is also the co-founder of Bird on-chain analytics protocol.

Issah Abubakari Samori has a background in Biomedical Engineering and currently works as an Artificial Intelligence Engineer. He focuses on building AI solutions for challenges in biomedicine and healthcare. Outside of building these AI solutions, he is also interested in the implications these solutions have on cybersecurity and biocybersecurity

Donna M. Schaeffer is a Professor in the School of Technology and Innovation at Marymount University. Dr. Schaeffer has published more 100 articles and papers, book chapters, and encyclopedia entries. She serves as Co-Chair of Transform Learning Working Group at the National Institute of Cybersecurity Education, and held leadership positions in professional associations.

Nkata Sekonya is a postgraduate student in Computer Science at the University of Johannesburg, SA. He received his BSc in Information Technology from the University of Johannesburg in 2021. His main research interests are digital forensics, critical infrastructure protection, and the internet of things.

Alan Shaffer is a Senior Lecturer and Associate Chair of Academics for the Computer Science Department at the Naval Postgraduate School. His research interests include cyber operations and security, computer systems modelling and verification, and data science. He has been involved with the IEEE Symposium on Security and Privacy as Organizing Committee Treasurer.

Siphesihle Sithungu is from Johannesburg, South Africa and holds an 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.

Dr Sidney Smith graduated from Towson University with a BS in Computer Science in 1990, a MS from TU in 2013, and a D.Sc. in 2019. He began his career with the US Army in 1990 and is currently a Computer Scientist for the Army Research Laboratory. He holds these professional certifications: CISSP, CISA, and CAP.

Simone Stephen is affiliated with the Old Dominion University School of Cybersecurity. Her areas of research interest currently include cyberbiosecurity, bioeconomics, mechanical engineering, aerospace engineering, space systems, cybersecurity, security awareness, environmental management, digitalization, education, and artificial intelligence.

Christelle Steyn, a student nearing the end of my masters studies at the University of Stellenbosch. I have an interest in cybersecurity and information systems and combined these fields with my interest in nature and wildlife conservation for my research.

Jake Street is a PhD student and Academic Associate and Nottingham Trent University, UK. His research is focused on creating solutions for protecting children online from social engineering attacks e.g. Online Grooming, as well as observing the additional factors that make these attacks complex to mitigate.

David Tileston is a cybersecurity engineer and exercise developer at Carnegie Mellon University's Software Engineering Institute (SEI). His experience at the SEI over the past 7 years includes dynamic defense research, cybersecurity curriculum development, cyber

exercise ranges, and infrastructure automation. David previously worked as a tool developer for 2 years as an Army Civilian.

Stephen Treacy, Anoop Sabu, Thomas Bond, Joseph O'Sullivan, Jack Sullivan, and Peter Sylvester are research collaborators based out of Cork University Business School, Ireland. This project was the result of a Masters' thesis developed from the MSc in Information Systems for Business Performance in conjunction with ForcePoint Cybersecurity.

Maureen S. Van Devender is currently pursuing a Ph.D. in Computing at the University of South Alabama with a research focus on cybersecurity risk assessment. Her publications include the topics of cybersecurity vulnerability of networked medical devices and privacy preserving techniques in data that contains personally identifiable information.

Namosha Veerasamy is a computer scientist in the field cyber security research and governance for over 15 years. She received her PhD from the University of Johannesburg. She is currently employed as a senior researcher at the Council for Scientific and Industrial Research (CSIR) in Pretoria. She also holds CISSP and CISM certifications.

Andreas Weinand received his Dipl.-Ing. degree (M.Sc) from the TU Kaiserslautern, Germany in 2015. He is currently pursuing the Dr.-Ing. degree (Ph.D) at the Institute for Wireless Communication and Navigation at the RPTU Kaiserslautern, Germany. His research interests are in the area of reliable and secure wireless systems for machine type communication (MTC).

Dana Beth Weinberg, PhD, is a Professor of Sociology and Data Analytic at Queens College-CUNY. She currently studies influence operations and the ways they leverage and weaponize narratives. She is the Director of the New War Research Consortium, a civil-military collaboration combining academic insights and practical solutions for national security policy.

Benjamin Yankson is an assistant professor and Cybersecurity practitioner with several years of experience in various technical leadership roles in information technology security within the healthcare and education industries. His research focuses on IoT Security, Privacy, Cybersecurity Threat and Risk Assessment, Security Auditing/Compliance, and Digital Forensics. In addition, he has served as a reviewer and TPC member of several conferences

Secure Cloud Migration Strategy (SCMS): A Safe Journey to the Cloud

Dalal N. Alharthi
University of Arizona, Tucson, AZ, United States

dalharthi@arizona.edu

Abstract: The state of cloud security is evolving. Many organizations are migrating their on-premises data centers to cloud networks at a rapid pace due to the benefits like cost-effectiveness, scalability, reliability, and flexibility. Yet, cloud environments also raise certain security concerns that may hinder their adoption. Cloud security threats may include data breaches/leaks, data loss, access management, insecure APIs, and misconfigured cloud storage. The security challenges associated with cloud computing have been widely studied in previous literature and different research groups. This paper conducted a systematic literature review and examined the research studies published between 2010 and 2023 within popular digital libraries. The paper then proposes a comprehensive Secure Cloud Migration Strategy (SCMS) that organizations can adopt to secure their cloud environment. The proposed SCMS consists of three main repeatable phases/processes, which are preparation; readiness and adoption; and testing. Among these phases, the author addresses tasks/projects from the different perspectives of the three cybersecurity teams, which are the blue team (defenders), the red team (attackers), and the yellow team (developers). This can be used by the Cloud Center of Excellence (CCoE) as a checklist that covers defending the cloud; attacking and abusing the cloud; and applying the security shift left concepts. In addition to that, the paper addresses the necessary cloud security documents/runbooks that should be developed and automated such as incident response runbook, disaster recovery planning, risk assessment methodology, and cloud security controls. Future research venues and open cloud security problems/issues were addressed throughout the paper. The ultimate goal is to support the development of a proper security system to an efficient cloud computing system to help harden organizations' cloud infrastructures and increase the cloud security awareness level, which is significant to national security. Furthermore, practitioners and researchers can use the proposed solutions to replicate and/or extend the proposed work.

Keywords: Cloud Migration, Cloud Transformation, Cloud Security, Cloud Strategy, Attacking Public Cloud, Defending Public Cloud

1. Introduction

Cloud solutions provide a powerful computing platform that enables organizations to migrate their on-premises applications and storage space as easy, flexible, and cost-effective solutions/services over the internet. Yet, cloud environments also raise certain security concerns that may hinder their adoption. Virtualization, storage, and networks, as mentioned by Ouda et al (2022), are the most serious security concerns in cloud computing. Hence, awareness of the latest innovations in cloud security technologies is necessary in order to mitigate the risks of these security challenges and provide security solutions to remain competitive (Hassan et al, 2019). Several research efforts have been made in evaluating challenges related to security faced by cloud computing environments, and a number of solutions to such problems have also been proposed. Despite the research efforts in the cloud security field, there are still some open research problems and challenges which are discussed in this paper.

As addressed by Ahmad et al (2020), migration to cloud environments needs careful planning, strong business case, credible migration strategy, and robust migration frameworks. Hence, this paper aims to develop a Secure Cloud Migration Strategy (SCMS). The proposed strategy addresses cloud migration phases, which are 1) preparation; 2) readiness and adoption; and 3) testing. Among these phases, cloud security tasks/projects are illustrated from the different perspectives of cloud defenders, attackers, and developers. Adopting the proposed SCMS may enable both academia and practitioners in the cloud computing community to get an overarching view of the strategy of the legacy application migration to public cloud environments in a secure manner. Additionally, the author addresses cloud security documents/runbooks that need to be developed, such as incident response runbook, disaster recovery planning, risk assessment methodology, and cloud security controls. Furthermore, the paper identifies some necessary cloud security challenges that have not yet been addressed by existing approaches, developing opportunities for further research endeavors.

The main contribution of this study can be summarized as follows:

- The paper summarizes current research efforts related to securing the cloud migration journey.

- The paper categorizes tasks/projects to secure the cloud migration journey from the different perspectives of defenders, attackers, and developers.
- The paper addresses cloud security documents/runbooks that organizations may develop and automate to secure the cloud migration journey.

The remainder of this paper is structured as follows. Section 2 provides the necessary background information on cloud providers in addition to the multi-cloud migration approach. Section 3 presents the related research efforts on cloud security. Section 4 addresses the research questions this paper aims to answer and describes the author's methodology for developing the strategy and identifying cloud security documents/runbooks. The proposed Secure Cloud Migration (SCM) Strategy is presented and illustrated in Section 5. Then, Section 6 highlights the documents/runbooks that organizations may develop and automate when migrating to the cloud. Section 7 addresses threats to the validity of the proposed solutions. The paper's conclusion and future work are mentioned in section 8.

2. Background

This section provides an overview of the big three cloud providers that organizations consider migrating to, which are Amazon Web Service (AWS), Microsoft Azure, and Google Cloud Platform (GCP). Additionally, it sheds some light on the multi-cloud approach.

AWS is the first-to-market launching in 2006. It has over 200 services as of February 2023. It is considered by many organizations to be the default choice for cloud migration due to its feature set, reliability, and growth. Azure is also gaining on AWS quickly due to its competitive pricing model. It has proven itself to be a viable alternative. Although GCP lags far behind them both, it too has grown significantly in market share. GCP is also a patron of open-source technology such as Kubernetes and TensorFlow. It also acquired some popular cloud platforms including Firebase and Stackdriver.

These large-scale cloud providers allow organizations to pay for whatever resources they need, enabling customers to increase or decrease the number of resources requested as needed. Yet, the traditional use of cloud services focused on the consumption of one provider is not valid anymore due to different shortcomings being the risk of vendor lock-in a critical (Alonso et al, 2023). Hence, the use of multiple cloud providers simultaneously is increasing to achieve higher service availability and reduce damage in the case of malicious insiders on a single Cloud Service Provider (CSP) (AlZain et al, 2012). Therefore, many organizations prefer to migrate their on-premises environments to a multi-cloud approach to leverage the best of all cloud worlds. Forrester surveyed 727 cloud decision-makers globally and found that 86% of respondents identified themselves as "multi-cloud" (McLellan, 2019). Multi-cloud can be defined as utilizing numerous cloud networks and services simultaneously (Hong et al, 2019). While migrating to a multi-cloud environment can help develop a more effective disaster recovery plan, it can also produce some challenges for cloud security professionals. As addressed by Kavitha and Radha (2022), a multi-cloud environment has the possibility of greater challenges for service quality, security, and privacy attacks.

3. Literature Review

In the existing literature, an increasing amount of research works addressed the security issues in cloud computing to find efficient security solutions. On the other hand, different research groups have studied and published cloud migration security mechanisms. Gonzalez et al (2012) addressed the most remarkable ones named NIST (National Institute of Standards and Technology), CSA (Cloud Security Alliance), and ENISA (European Network and Information Security Agency). This section focuses on the related research efforts in light of this research.

Cloud computing, a new paradigm of flexible and cost-effective computing, is being explored as a possible solution to make enterprise IT sustainable. In that context, Ahmad, Naveed, and Hoda (2018) highlighted some of the procedures for cloud migration planning and execution, such as architecture recovery, dependency checks, provisioning, pilot migration, data migration, application migration, integration, and validations. Among these procedures, there are key processes concepts, as addressed by Fahmideh et al (2019), that need to be incorporated into a typical process of legacy system migration to the cloud including the following: analyze migration requirements, define a plan, recover legacy system knowledge, choose cloud platform/provider, design cloud solution, identify incompatibilities, make system stateless, decouple system components, replicate system components, make mock migration, use logging, resolve licensing issues, develop integrators, deploy

system component, enable elasticity, encrypt the database, handle transient faults, isolate tenant, encrypt/decrypt messages, obfuscate codes, reconfigure the network, synchronize/replicate system components, communicate a-synchronous, and test system.

Due to its dynamic abstraction and scalability, applications and data outsourced to the cloud have unlimited security boundaries and infrastructure. Another primary security concern surrounding the adoption of cloud computing is its multi-tenancy nature and sharing of virtualized resources (Abdulsalam and Hedabou, 2022). To secure the systems migration to the cloud environments, it is vital to be aware of crucial aspects of security. El Kafhali, El Mir, and Hanini (2022) and Masadeh, AlShrouf, and Kumar (2023) identified some of these security challenges and issues such as the lack of control over the Cloud Datacenters, weak integration of security into the service development process, data breaches, data leakage/loss, account or service traffic hijacking, insecure interfaces/APIs, Denial of Service (DoS), malicious insiders, insufficient due diligence, shared technology vulnerabilities, loss of governance, lock-in, insecure or incomplete data deletion, and the availability chain.

4. Research Questions and Methodology

This section describes the research questions this research aims to answer, followed by the research methodology to develop the strategy and identify cloud security documents/runbooks.

- RQ1: What are the main phases of migrating on-premises applications and services to the cloud?
- RQ2: What are the key tasks/projects to secure the cloud migration journey from the different perspectives of defenders, attackers, and developers?
- RQ3: What are the necessary cloud security documents/runbooks that organizations may develop and automate to secure the cloud migration journey?

To answer these questions, the author followed a Systematic Literature Review (SLR) technique as recommended by Okoli and Schabram (2010). The study conducted a literature review of the most recent journal and conference papers that contained the following keywords and strings in their titles “Cloud Migration”, “Cloud Security”, and/or “Cloud Strategy”. For each paper, the author extracted security challenges from both offensive and defensive security perspectives. This SLR examined the research studies published between 2010 and 2023 within popular digital libraries such as IEEE Xplore, Springer, ACM, Google Scholars, and Science Direct. These digital libraries are the primary source of publications on topics from both computer science and cybersecurity domains.

5. Secure Cloud Migration Strategy (SCMS)

By investigating the existing literature, the author classifies the Secure Cloud Migration Strategy (SCMS) into three main repeatable phases/processes for iterative cloud transformation, which are 1) Preparation; 2) Readiness and Adoption; and 3) Testing. Among these processes, proposed cloud security projects are addressed from the different perspectives of defenders, attackers, and developers. More details on the proposed SCMS are addressed below and illustrated in Figure 1.

5.1 Preparation

This phase includes tasks such as building the Cloud Center of Excellence (CCoE), designing the migration architecture, and developing the RACI Chart. Building the CCoE can be one of the first steps in the cloud migration journey. This should include the following teams: Infrastructure, Cloud Security, Developers, Contractors, and anyone who is working on the cloud migration project. Another task to be completed in this phase is developing the Cloud Migration Architecture, which should take into consideration Business Units, Core Accounts, Firewalls, and the Transit Gateway and/or VPC Peering. Moreover, the team should invest in building the cloud migration RACI Chart. The acronym RACI stands for responsible, accountable, consulted, and informed. It is, as described by Araújo (2021), a clear matrix chart used to attribute responsibilities and duties per each job or decision in a project by all of the team members.

5.2 Readiness and Adoption

In this phase, both defenders (blue team) and developers (yellow team) may work together to ensure the security of the cloud environment. Here are more details on what tasks/projects can be included in this phase. From the defenders’ perspectives, projects/tasks may start with preparing the security toolkits such as firewalls, Single Sign On (SSO), password management, antivirus, SIEM solution, vulnerability assessment, and compliance

tools. Integrating these tools is essential i.e., the vulnerability assessment tool may be integrated with the compliance tool to provide a better understanding of the current compliance status of the cloud environment. Other tasks/projects may include Identity and Access Management (IAM), least privileges determination, access control, secret management, data encryption, network security, defense-in-depth, continuous monitoring, and cloud compliance to the relevant industry standards and regulations. From the developers' perspective, it is vital to adopt the Security Shift Left concept to shift security to be in the early stages of the Continuous Integration and Continuous Deployment (CICD) pipeline of the software development cycle. Gonzalez, Perez, and Mirakhorli (2021) highlighted seven pain points that developers may experience as they learn to adapt their existing skills to the security tests, which are mocking security components, interacting with security services/APIs, creating security fixtures, bypassing access controls for protected resources, designing authentication flows, creating HTTP request objects, and configuring systems for test environments.

5.3 Testing

Ease of using the cloud may cause security flaws and due to this reason, security issues of cloud computing should be examined in-depth (Yurtseven and Bagriyanik, 2020). Therefore, security testing is fundamental to identifying existing security issues and is particularly powerful when carried out by means of penetration testing (pentesting) and vulnerability assessment techniques. Casola et al (2018) proposed a methodology that allows easy carrying out of a coarse-grained security evaluation of a cloud application by automating the set-up and execution of penetration tests. The methodology relies on the knowledge of the application architecture and on the availability of a catalog including security-related data collected from multiple sources and properly correlated. The methodology started with the preparation phase in which a risk assessment should be conducted, followed by the scanning phase to identify weaknesses and vulnerabilities, then the pentesting phase to configure and execute the attack. The tasks/projects identified here are the responsibility of the red team.

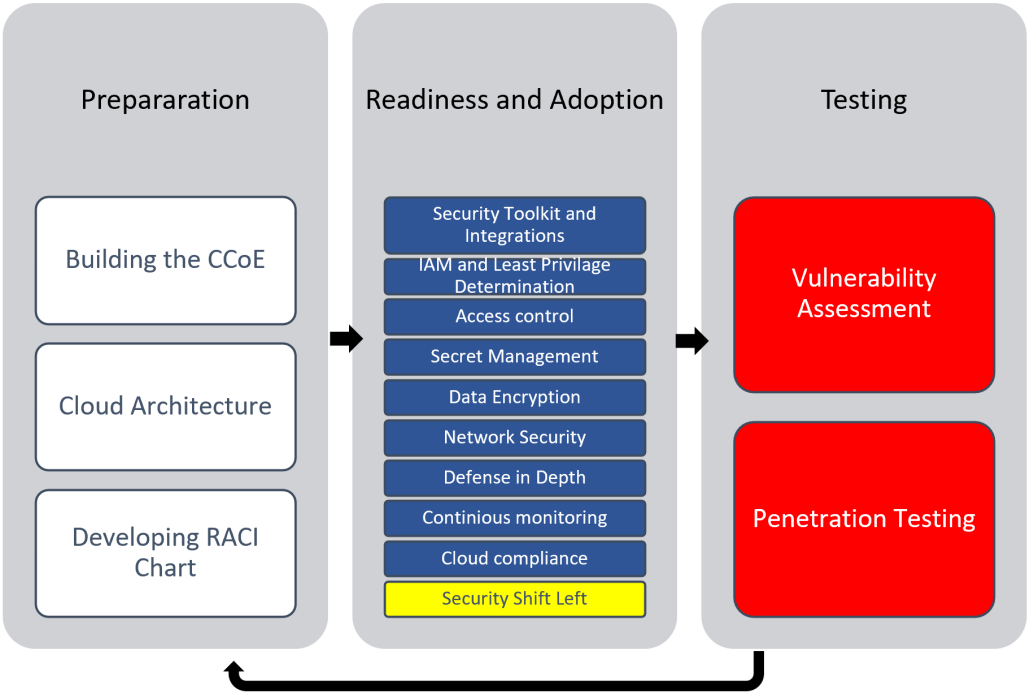


Figure 1: The proposed Secure Cloud Migration Strategy (SCMS)

6. Secure Cloud Migration Documents/Runbooks

Among the three phases addressed by the SCMS in Section 6, it is essential to develop and automate cloud security documents/runbooks to ensure a secure journey to cloud transformation. This section addresses some of these documents/runbooks, such as the Incident Response (IR) Runbook, Risk Assessment Methodology, Disaster Recovery Planning, and Cloud Security Controls.

6.1 The Incident Response (IR) Runbook

This should consider incident response steps addressed by Grobauer and Schreck (2010), which include containment, eradication, and recovery. Upon developing the document, it should be tested and automated. The IR Runbook may also contain the contact information of the teams who will be responsible for responding to cloud security incidents. Additionally, different cyberattack scenarios should be illustrated, and the appropriate response to each of the scenarios shall be addressed in detail. NIST outlined these response steps in (Cichonski et al, 2012) to be gathering evidence; containing and then eradicating the incident; recovering from the incident; and conducting post-incident activities, including post-mortem and feedback processes.

6.2 Disaster Recovery Planning

As the name implies, this document shall contain a recovery plan in place in case of data loss or breaches. The primary concern for this document is finding ways to ensure that the process of data backup and recovery is effective in providing high data availability, flexibility, and reliability at a reasonable cost. As addressed by Alshammari et al (2017), the majority of providers of cloud services enable their customers to rapidly recover from disasters with a minimum of disruption via the use of a geographically distributed data backup and redundancy model.

6.3 Risk Assessment Methodology

This document shall identify potential security risks and data privacy issues associated with migrating to the cloud. According to Ali et al (2020), unique elements of risk assessment in cloud computing may relate to the operational security and monitoring of cloud services. Security assessments are important mechanisms for risk mitigation from cloud security breaches.

6.4 Cloud Security Controls

This document can serve as guidance for developers to address cloud security controls that should be taken into consideration when migrating on-premises applications/services to the cloud. Developing this document may help the implementation of SecDevOps, which has recently appeared in the researchers and developers communities, but the management of security in a DevOps life cycle is not a straightforward task. Hence, adopting Security-by-Design through developing and implementing the Cloud Security Controls document is vital. The Security-by-Design approach, as defined by Casola et al (2020) and Santos, Tarrit, and Mirakhorli (2017), is designing the software from the foundation to be secure. At this aim, the alternate security tactics and patterns are first thought and, among them, the best are selected and enforced by the application designer, and then used as guiding principles for developers.

7. Threats to Validity

This section discusses the threats to the validity of the proposed solutions and the author's steps to minimize those threats. To develop the SCMS and the associated tasks/projects to secure cloud transformation in addition to identifying cloud security documents/runbooks, the author mainly relied on a comprehensive literature review. As with any such review, some significant references could have gone unnoticed. To minimize this threat, the author examined papers with “Cloud Migration”, “Cloud Security”, and “Cloud Transformation” keywords and strings in the title and read their abstracts, introductions, and conclusions.

8. Conclusion and Future Work

Being a relatively newer technology, cloud computing comes with its own array of security challenges and new perspectives for attacks to happen. Despite the rigorous efforts of the cloud security research community, there are still many open issues and challenges to the security of the cloud environment. Hence, this paper proposed a secure path to cloud transformation through a Secure Cloud Migration Strategy (SCMS). Additionally, necessary cloud security documents/runbooks were identified in this paper. Both contributions were developed based on a systematic literature review on the legacy of securing the cloud migration journey.

The next step is to analyze and test the effectiveness of the proposed SCMS and cloud security documents/runbooks by implementing them through an experimental study involving migrating applications/services to cloud environments. As another venue of future work, the author aims to develop a dynamic-optimizing approach for a secure and relaxed privilege management system in the cloud. The term

'secure and relaxed' refers to usable security which aims to make sure that security products and processes are usable by those who need them.

References

Abdulsalam, Y.S. and Hedabou, M., 2022. Security and privacy in cloud computing: technical review. *Future Internet, 14*(1), p.11.

Ahmad, N., Qamar, S., Khan, N., Naim, A., Hussain, M.R., Naveed, Q.N. and Mahmood, M.R., 2020. Cloud computing trends and cloud migration tuple. *Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019,* pp.737-745.

Ahmad, N., Naveed, Q.N. and Hoda, N., 2018, November. Strategy and procedures for Migration to the Cloud Computing. In *2018 IEEE 5th international conference on engineering technologies and applied sciences (ICETAS)* (pp. 1-5). IEEE.

Ali, O., Shrestha, A., Chatfield, A. and Murray, P., 2020. Assessing information security risks in the cloud: A case study of Australian local government authorities. *Government Information Quarterly, 37*(1), p.101419.

Alonso, J., Orue-Echevarria, L., Casola, V., Torre, A.I., Huarte, M., Osaba, E. and Lobo, J.L., 2023. Understanding the challenges and novel architectural models of multi-cloud native applications–a systematic literature review. *Journal of Cloud Computing, 12*(1), pp.1-34.

Alshammari, M.M., Alwan, A.A., Nordin, A. and Al-Shaikhli, I.F., 2017, November. Disaster recovery in single-cloud and multi-cloud environments: Issues and challenges. In *2017 4th IEEE international conference on engineering technologies and applied sciences (ICETAS)* (pp. 1-7). IEEE.

AlZain, M.A., Pardede, E., Soh, B. and Thom, J.A., 2012, January. Cloud computing security: from single to multi-clouds. In *2012 45th Hawaii International Conference on System Sciences* (pp. 5490-5499). IEEE.

Araújo, T.M.R.P.D., 2021. Cloud Technology Selection: *A structured framework for decision making (Doctoral dissertation).*

Casola, V., De Benedictis, A., Rak, M. and Salzillo, G., 2020. A cloud SecDevOps methodology: from design to testing. In *Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9–11, 2020, Proceedings 13* (pp. 317-331). Springer International Publishing.

Casola, V., De Benedictis, A., Rak, M. and Villano, U., 2018, June. Towards automated penetration testing for cloud applications. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 24-29). IEEE.

Cichonski, P., Millar, T., Grance, T. and Scarfone, K., 2012. Computer security incident handling guide. *NIST Special Publication, 800*(61), pp.1-147.

El Kafhali, S., El Mir, I. and Hanini, M., 2022. Security threats, defense mechanisms, challenges, and future directions in cloud computing. *Archives of Computational Methods in Engineering, 29*(1), pp.223-246.

Fahmideh, M., Daneshgar, F., Rabhi, F. and Beydoun, G., 2019. A generic cloud migration process model. *European Journal of Information Systems,* 28(3), pp.233-255.

Gonzalez, D., Perez, P.P. and Mirakhorli, M., 2021, October. Barriers to Shift-Left Security: The Unique Pain Points of Writing Automated Tests Involving Security Controls. In *Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (pp. 1-12).

Gonzalez, N., Miers, C., Redigolo, F., Simplicio, M., Carvalho, T., Näslund, M. and Pourzandi, M., 2012. A quantitative analysis of current security concerns and solutions for cloud computing. *Journal of Cloud Computing: Advances, Systems and Applications, 1*, pp.1-18.

Grobauer, B. and Schreck, T., 2010, October. Towards incident handling in the cloud: challenges and approaches. In *Proceedings of the 2010 ACM workshop on Cloud computing security workshop* (pp. 77-86).

Hassan, W., Chou, T.S., Li, X., Appiah-Kubi, P. and Tamer, O., 2019. Latest trends, challenges and solutions in security in the era of cloud computing and software defined networks. *Int J Inf & Commun Technol ISSN, 2252*(8776), p.8776.

Hong, J., Dreibholz, T., Schenkel, J.A. and Hu, J.A., 2019. An overview of multi-cloud computing. In *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33* (pp. 1055-1068). Springer International Publishing.

Kavitha, M.G. and Radha, D., 2022. Quality, Security Issues, and Challenges in Multi-cloud Environment: A Comprehensive Review. *Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases,* pp.269-285.

Masadeh, S.R., AlShrouf, F.M. and Kumar, A.S., 2023. Concerns from Cloud Security Issues: Challenges and Open Problems. *International Journal, 12*(1).

McLellan, C., 2019. Multicloud: Everything you need to know about the biggest trend in cloud computing. *ZDnet, Jul.*

Okoli, C. and Schabram, K., 2010. A guide to conducting a systematic literature review of information systems research.

Ouda, A.J., Yousif, A.N., Hasan, A.S., Ibrahim, H.M. and Shyaa, M.A., 2022. The impact of cloud computing on network security and the risk for organization behaviors. *Webology, 19(1)*, pp.195-206.

Santos, J.C., Tarrit, K. and Mirakhorli, M., 2017, April. A catalog of security architecture weaknesses. In *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)* (pp. 220-223). IEEE.

Yurtseven, I. and Bagriyanik, S., 2020, October. A review of penetration testing and vulnerability assessment in cloud environment. In *2020 Turkish National Software Engineering Symposium (UYMS)* (pp. 1-6). IEEE.

Cyber-Physical Attack Using High Power RF in Havana, Cuba

Allyffazzkkamn Argudo and Ghislaine Nasibu Mwana
Marymount University, Alexandria Virginia, USA

ara47724@marymount.edu
gmn64459@marymount.edu

Abstract: Global communications have relied heavily on fiber-optic cables. Satellite communications have become the norm for areas that are not highly populated or easily reached by conventional wires. The availability and use of satellite communications systems have had a limited market; with this limitation, the need for open development of each vendor's software never gained traction. This led to weaknesses not being discovered until a flaw was made public in a systems breach that affected the end users. Wireless communications that rely on microwaves lead to the use of malware to provide an override to normal operations of the equipment. As the issue of Stuxnet would allow for the override of the safety settings of satellite terminals, the malware could even be deployed remotely through a system update. The weaponization of such systems is now a point of concern. Previous studies have shown that naval satellite communications systems can be a weapon by removing the software power limitations (McKay, 2021). As satellite systems on ships can be larger due to being fixed, transmission power generation may be more significant. This study considers the land-based mobile systems deployed against any target and quickly dismantled and removed. Can land-based mobile satellite communication systems affected by Stuxnet or manually altered yield the same power output, and at what range will it affect human tissue? This is believed to be the case with the ongoing investigation. The survey results showed that mobile equipment could be dangerous to human tissue at distances easily achieved by mobile terminals, especially if the target is a fixed location like a building with large glass windows that R.F. power can penetrate easily.

Keywords: Replication Study, Cyber-Attack, SATCOM, VSAT, BGAN, HIRF

1. Introduction

Maritime trade is critical to the global economy. Globally, 90% of the world's traded goods move thru waves (Arampatzis, 2021). Cyber threats in the Maritime industry can seriously affect the national economy due to the digitalization of the infrastructure, which makes it vulnerable. In the U.S., it accounts for a third of the country's economy. The increased dependence on the internet equates to increased vulnerability. Maritime communications between cruise ships or cargo shipping depend mainly on non-terrestrial communication. Terrestrial communications rely on satellite communications to provide access to more remote areas where communications are required. Military, government agencies, and commercial entities must have contacts through reliable means as some of the market locations may not have steady and reliable traditional communication networks.

SATCOM network plays an essential role in global telecommunications systems. They are the roots of various services used on vessels to guarantee safety and security (Caprolu et al., 2020). In the maritime industry, SATCOM is seen as a global WIFI network. "Its core is linked together through network nodes just as a regular office network would be; except, in this case, R.F. energy is used in place of cables, which propagates data payloads throughout great distances in air and space on the backbone of SATCOM devices, satellites and ground nodes" (McKay, 2021). SATCOM network provides reliable communications to cruise ships and maritime vessels. Degraded communication is the major challenge in satellite communication systems. Vulnerabilities, in this case, include design flaws that allow access to the SATCOM by letting remote attackers intercept, manipulate, block, and in some cases, take complete control.

Problem Statement: Compromised satellite communication terminals, whether remotely accessed due to software vulnerabilities or controlled on-site due to a lack of physical safety mechanisms, can yield High-Intensity Radio Fields (HIRF). A by-product of exceeding safety limits can cause harm to human tissue in direct relation to the proximity of the source.

2. Literature review

2.1 Introduction

This research utilizes similar parameters as those provided by McKay (2021). The land-based study of how HIRF is achieved thru smaller mobile terminals that can easily be concealed and aimed at targets of interest was added.

2.2 Defining Cyber-physical attacks

"The rocket worked perfectly except that it landed on the wrong planet" – Werner von Braun stated this when his invention was turned into a weapon of war. As with any new technological development thought to aid humanity, the same concept can then be used to harm. Cyber-physical attacks are such a new development that many possible outcomes have yet to be documented. Cyber-physical attacks have become part of real life; what once was a science fiction idea is now a daily concern as the internet of things (IoT) has grown. The exploitation weaknesses within a physical system connected to the internet are what cybercriminals focus on with the intent to alter or take over the functionality of a physical system. For the scope of this study, cyber-physical attacks will focus on high radiofrequency radiation and the possibility of being able to cause harm to biological specimens through the alteration of energy applied to a system from naval equipment and land-based equipment.

2.3 Satellite Hacking

In his study "Satellite network hacking & security analysis," published in 2016, Adam Hudaib discussed satellite hacking in-depth. One central point that Hudaib made is that most satellites are old and critical infrastructure, and these lapses make them vulnerable to hacking. Huda states that "older satellites might be more vulnerable as defenses may be less stringent, security flaws can be exposed over time, and there are limited means to patch those flaws once it is in orbit" (Hudaib, 2016, p34). Huda has broken-down satellite hacking into four main types: (1) Jamming: refers to flooding or overpowering a signal, transmitter, or receiver so that the legitimate transmission cannot reach its destination. (2) Eavesdropping: allows a hacker to see and hear what is being transmitted. (3) Hijacking: the unauthorized use of a satellite for transmission or seizing control of a signal and replacing it with another. (4) Control: taking control of part or all of the TT&C ground station (Hudaib, 2016, p38). "Last Call for SATCOM Security" springboards off a 2014 white paper, "A Wake-up Call for SATCOM Security. Santamarta (2014) provides insight into how satellite systems for all locations have vulnerabilities that were not addressed, as cybersecurity was a field being developed. Most non-state attackers were not focused on utilizing SATCOM systems to conduct any attack, mainly due to the volatile stability. SATCOM systems are turned off when not in use since satellite systems are charged per hour used, or fuel consumption to power generators would not be efficient when a communications system is not required. The studies' authors provided insight into the significant vulnerabilities of aerial, maritime, military/land, and space systems. The tables below show the information matrix that the authors of the study compiled:

Table 1: Attack Vector Identified

Industry	Security Risk	Flight Safety Risk	RF Risk	Likelihood	Attack Vector
Aviation	Yes	No*	No*	Medium	Remote
Maritime	Yes	N/A	Yes	High	Remote
Military	Yes	N/A	No	Medium	Remote

*Based on input received from the Aviation industry through the A-ISAC and our own research

Table 2: Vulnerability Outcome

Industry	Threat
Aviation	• Ability to disrupt, intercept or modify non-safety communications such as In-Flight WiFi* • Ability to attach crew and passenger devices. • Ability to manipulate SATCOM antenna positioning and transmissions.
Maritime	• Ability to disrupt, intercept or modify onboard satellite communications. • Ability to attack crew's devices. • Ability to control SATCOM antenna positioning and transmissions. • Ability to perform cyber-physical attacks using HIRF.
Military	• Ability to pinpoint the location of military units. • Ability to disrupt, intercept or modify satellite communications. • Ability to perform cyber-physical attacks using HIRF.
Space	• Ability to disrupt satellite transponders.

* Typically pilot and co-pilot do not use it. In-Flight WiFi is normally used by flight attendants for PAX and PCI transactions. * Configurations may vary the impact. Santamarta 2018, p2. Appendix I

Probable capabilities of state-level entities were not provided. The study showed many vulnerabilities in many commercial systems only. The authors urged the parties whose equipment was studied to fix the deficiencies. However, as most satellite systems are costly to develop, most companies that have spent many resources in developing and deploying these systems have proprietary protocols that are closed source, making them undocumented. While there is something to be said about security thru obscurity, once an adversary can figure out a system, the ability to react or have other entities analyze the system's weaknesses is hindered by the same security thru obscurity. One example of cooperation was INMARSAT when it cooperated with a third party to solve a problem with the marine communications system. A back door was discovered that allowed for remote installation of the unauthorized firmware update (Caprolou, et al., 2020, p4). This kind of action yields the inevitable possibility of being able to remotely program a VSAT dish and other programmable antennas and overriding pre-programmed power transmission and position of antennas. While many systems utilized in maritime and ground operations may have newer hardware as of 2016, there are 1046 satellites in space; most have hardware and programming relevant to the technology of the 1980s. The issue of having ground equipment that meets today's communications standards while still being able to translate the information to older systems creates avenues of attack. About 1.4 million VSAT terminals are connected to satellites at any given time. That sheer quantity of terminals establishes a target of opportunity. Most maritime terminals that are part of this replication study have had known issues. The land terminals need not have any problems as a nefarious actor can purchase a system and aim it at a target manually while overriding safety protocols of the equipment thru physical manipulation or software manipulation.

3. Communications equipment

For this replication study, all BGAN and VSAT terminals are considered relevant. A summary of observed cyber issues for many systems has been provided in the tables 3 and 4 below. By adding newer land systems that have not been tested to have cybersecurity issues, this study assumes that software can be remotely accessed, along with in-person adjustments.

Table 3: Hawkeye 2019



ANTENNA			
Aperture	1.3m parabolic		
Optics	Center fed	Auto-Acquire	
Elevation	Manual	5 to 90 degrees	
Azimuth	90 degrees	120 degrees	
LEO/MEO Capable	N/A	Yes	
PACKED SYSTEM WEIGHT (950MP W/AC POWER SUPPLY			
Weight	Two cases, <84lbs		
Dimensions	All cases: <70 linear in		
RF PERFORMANCE			
RF BAND	X	Ku	Ka
Transmit (GHz)	7.90 - 8.40	13.75 - 14.50	29.0 - 31.0
Receive (GHz)	7.25 - 7.75	10.95 - 12.75	19.20 - 21.20
Polarization	Circular	Linear	Circular
G/T	16.7 dB / K	20.5 dB / K	21.8 dB / K
EIRP (Standard Power)	51.7 dBW	54.8 dBW	56.6 dBW
EIRP (High Power)	54.7 dBW	58.2 dBW	59.9 dBW

MODEM INTEGRATION			
Supported Modems	iDirect 950mp (Evolution and Velocity)		
	ViaSat CBM-400 (Linkway and EBEM)		
	L3Harris MPM-2500 (NCW)		
	External Modem		
Future Modems	iDirect450 mp		
	Comtech 5650C		
	L3Harris Protected Anti-Jam Tactical Satcom (PATS)		

Table 4: Star Win, 2021



Technical Specification:	
Type of antenna panel	Flat array antenna
Frequency Range	Tx 13.75 ~ 14.50 GHz Rx 10.95 ~ 12.75 GHz
Polarization	Linear polarization
VSWR	≤ -15dB
Cross Polarization Isolation	≥ 35dB
Tx/Rx Polarization Isolation	≥ 85dB
Satellite Searching Mode	Satellite parameter presetting GPS/Inclinometer guiding
Tracking Time	<120 S
BUC mounted panel	2\3\4\8\16W
Antenna Power Resistance	≥ 100W
Power Supply	AC 90-240V (DC optional)
Antenna Panel Size / Weight	540 x 570 x 34mm / 5Kg x 2
Total Antenna Weight	24KG

4. Harm to Biological Tissue

Published scientific literature on possible biological harm from animal or human exposure to R.F. energy. For many years now, it has been known that radio energy of sufficient intensity can produce heating in substances with finite conductivity, such as biological tissue. High R.F. levels can harm biological tissue due to the human body's inability to cope with or dissipate the excessive heat it generates. In other words, the specific physical reactions associated with R.F. field exposure are generally related to the energy absorption rate or the intensity of currents and internal electric fields. The FCC has conducted studies that show biological tissue's specific absorption rate of radio wave energy. The Specific Absorption Rate (SAR) is usually expressed as watts/kilogram. The safe rate of abortion for a human body is considered to be 0.04 W/kg OET, 56 [20] within a given period, reflected as 30 minutes. Table 5 below has specific information on absorption rates at different frequencies (Rooker, 2008) within a given period, reflected as 30 minutes. The table below has detailed information on different absorption rates at different frequencies. Some examples of physical harm come from the attacked that happened to the Indian army with the utilization of microwave pulses (Marlatt, 2005). Furthermore the National Academies of Science has concluded that the attack on the embassy in La Habana Cuba utilized radio frequency waves (Santamarta, 2014).

Table 5: Limits for Localized Exposure

Table 2. FCC Limits for Localized (Partial-body) Exposure	
Specific Absorption Rate (SAR)	
Occupational / Controlled Exposure	General Uncontrolled / Exposure
(100 kHz - 6 GHz)	(100 kHz -6 GHz)
< 0.4 W/kg whole-body	< 0.08 W/kg whole-body
<8 W/kg partial-body	< 1.6 W/kg partial-body

Table 6: Limits for Maximum Permissible Exposure

Table 1. FCC Limits for Maximum Permissible Exposure (MPE) (A) Limits for Occupational / Controlled Exposure				
Frequency Range (MHz)	**Electric Field Strength (E) (V/m)**	**Magnetic Field Strength (H) (A/m)**	**Power Density (S) (mW/cm^3)**	**Averaging Time $\|E\|^2$, $\|H\|^2$ or S (minutes)**
0.3 - 3.0	614	1.63	(100)*	6
3.0 - 30	1842/r	4.89/r	($900/f^2$)*	6
30 - 300	61.4	0.163	1.0	6
300 - 1500	-----	----	f/300	6
1500 - 100,000	-----	----	5	6
(B) Limits for General Population / Uncontrolled Exposure				
Frequency Range (MHz)	**Electric Field Strength (E) (V/m)**	**Magnetic Field Strength (H) (A/m)**	**Power Density (S) (mW/cm^3)**	**Averaging Time $\|E\|^2$, $\|H\|^2$ or S (minutes)**
0.3 - 1.34	614	1.63	(100)*	30
1.34 - 30	824/r	2.19/r	($180/f^2$)*	30
30 - 300	27.5	0.073	0.2	30
300 - 1500	----	----	f / 1500	30
1500 - 100,000	----	----	1.0	30

f = frequency in MHz

*Plane-wave equivalent power density

Note 1: Occupational / controlled limits apply in situations in which person are expose as a consequence of their employment provided those persons are fully aware of the potential for exposure and can exercise control over their exposure. Limits for occupational / controlled exposure also apply in situations when an individual is transient through a location where occupational / controlled limits apply provided he or she is made aware of the potential exposure.

Note 2: General population / uncontrolled limits exposures apply in situations in which the general public may be exposed, or in which persons that are exposed as a consequence of their employment may not be fully aware of the potential for exposure or cannot exercise over their exposure.

5. Deployment

The studied systems have already been placed within ships, aircraft, and land. Many commercial entities have heavily relied on maritime systems for communications to shore anywhere in the world. The Intellian GX60TM is a system installed in large cruise ships with antennas capable of handling large data transmissions. Intellian GX 100TM is another approach utilized in maritime activities such as cruise ships. The capacity for voice and data is more significant and can be used with different size antennas. The L3Harris Hawkeye 4TM lite is a system that can be rapidly deployed on land and provided with concealment inside larger structures that allow radio waves to travel through. Starwin flat panel system, meant for fast deployment and easy setup and tear down due to the shape. Only the tripod has to be removed for transport. The body alone allows for easier concealment from buildings and can even be placed within a vehicle to make it mobile and concealable. One major issue with these systems is the antenna size and distance for the field of effect. The smaller the antenna, the closer the target

needs to reach the necessary levels to cause harm due to the antenna's limitations of energy concentration and focal point.

6. Research methodology

Our study's methodology tried to reproduce McKay (2021) wherever possible. We used the same measures and identical treatment statistics as the original author's study, a quantitative observational simulation depicted by formulaic-driven simulations. The concern of the study is the possibility of thermic damage to a human or other biological tissue fluids.

6.1 Data Collection Set

The data collection will focus on the maritime systems with end frequencies due to the power needed to produce such frequencies. The GX60 system will provide the methodology for replicability. GX 100 is limited by propagation's possible power output at 10 watts. V80G system is built with varying power ranges; however, it provides this ability at lower frequencies. Yet the variable wattage of 4W, 6W, 8W, and 16W. For this study, the power level will be raised to 100 watts as it can override systems to this output level. To include newer land-based systems than can easily be assembled, hidden, and manually aimed at a specific target. The propagation of radiofrequency radiation can be calculated by the Far-Field Region formula as described in the FCC OET 65 Bulletin. The formulas below would provide the distance from which each antenna may propagate the radiofrequency radiation.

$$S = \frac{PG}{4\pi R^2}$$

where: S = power density (in appropriate units, mW/cm^2)
P = power input to the antenna (in appropriate units, e.g., mW)
G = power gain of the antenna in the direction of interest relative to an isotropic radiator
R = distance to the center radiation of the antenna (appropriate units, e.g., cm)

or:

$$S = \frac{EIRP}{4\pi R^2}$$

Where: EIRP = equivalent (or effective) isotropically radiated power

Another factor to consider comes from the second hypothesis. The ability to calculate radiofrequency radiation exposure from multiple frequencies. As satellite communication systems can broadcast more than one frequency. The formulas come from the International EMF project, which has studied this health issue and environmental factors of static radiofrequency radiation.

SIMULTANEOUS EXPOSURE TO MULTIPLE FREQUENCY FIELDS

It is important to determine whether in situations of these exposures are additive in their effects. Additivity and electrical stimulation, and the basic restrictions below should be met. The formulae below apply to relevant frequencies under practical exposure situations. For electrical stimulation, relevant for frequencies up to 10 MHz, induced current densities should be added to

$$\sum_{i=1\ Hz}^{10\ MHz} \frac{J_i}{J_{L,i}} \leq 1,$$

For thermal effects, relevant at 100kHz, SAR and power density values should be added according to:

$$\sum_{i=100\ kHz}^{10\ GHz} \frac{SAR_i}{SAR_L} + \sum_{i>10\ GHz}^{300\ GHz} \frac{S_i}{S_L} \leq 1$$

where:

J_i = the current density induced at frequency i;

$J_{L,I}$ = the induced current density restriction at frequency i as a given in Table 4;
SAR_i = the SAR caused by exposure at frequency i;
SAR_L = the SAR limit given in Table 4;
S_L = the power density limit given in Table 5; and
S_i = the power density at frequency i.

Calculating the near and far-field of an antenna, the formulas that the FCC published on OET bulletin 65 are shown in the table below.

For this study, the near-field energy is not being measured as it would have the biological matter close to the antenna. The far field is of interest as it is outside what most persons would consider a safe range, making it the ideal place to deliver an attack to those not suspecting. While distance calculations show that the area with the highest effect is considered near-field, as shown in the formula below OET, 65 (White House Trump Administration, 2020). For this study, the far-field effect will be observed and measured.

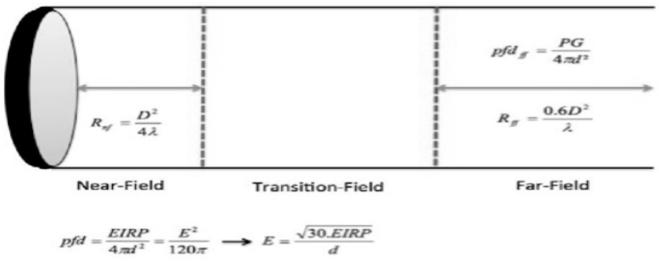


Figure 1: Energy Field

6.2 Tools to measure the results of the study

EIRP calculator (Effective Isotropic Radiated Power) is the measured radiated power of an antenna in a specific direction. It is also called Equivalent Isotropic Radiated Power. It is the output power when the Antenna concentrates a signal into a smaller area. The EIRP can consider the losses in the transmission line and connectors and includes the antenna's gain represented in dB. Entering the transmitted power, cable loss, and antenna gain to calculate the EIRP (Effective Isotropic Radiated Power) R.F., 2018 (National Academies of Sciences, 2020).

Antenna Factor and Gain Calculator

The antenna factor provides the ratio of the incident Electromagnetic Field to the output voltage from the antenna and the output connector. The Gain factor is the ratio of the signal received or transmitted by a given antenna compared to an isotropic or dipole antenna. Antenna gain can only be achieved by making an antenna directional, with better performance in one direction than in others Systems, 2018 (Marlatt, 2005).

7. Findings / Results

The values have been converted from voltage to watts as calculations for radio waves are compiled in wattage.

The study considers the average size of a human and weight to be about 150 lbs, based on the calculations conducted by the FCC. It should be noted that specific body parts /organs, due to their more direct exposure to the environment and greater volume of water per cm2, may be affected faster than organs further inside the body. The conductivity of the human body is set at 0.6 s.m, and parts of the body can conduct at a higher rate, such as the cerebral spinal fluid at 1.79 s/m Baumann (1997). Calculating the Specific Absorption Rate (SAR), the density of a 150lbs person will be set at the standard 1010 kg/m3. The electric field calculation from EIRP will be multiplied by 0.6 s/m divided by the density of 1010 kg/m3.

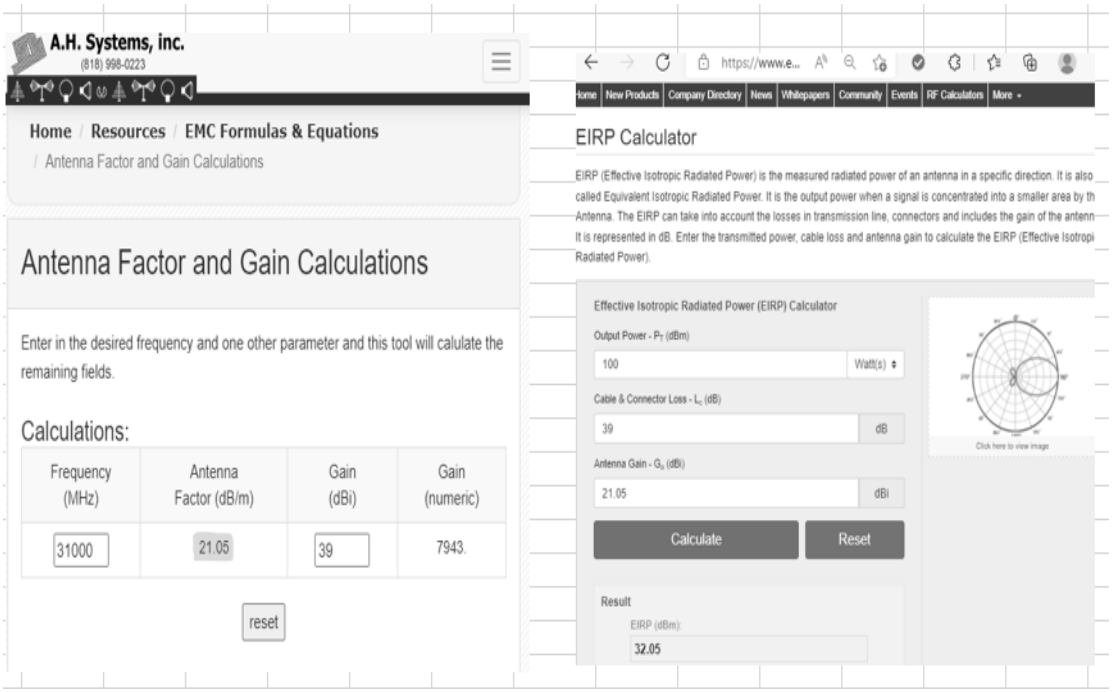


Figure 2: Calculations for L3Harris Hawk 4 Lite

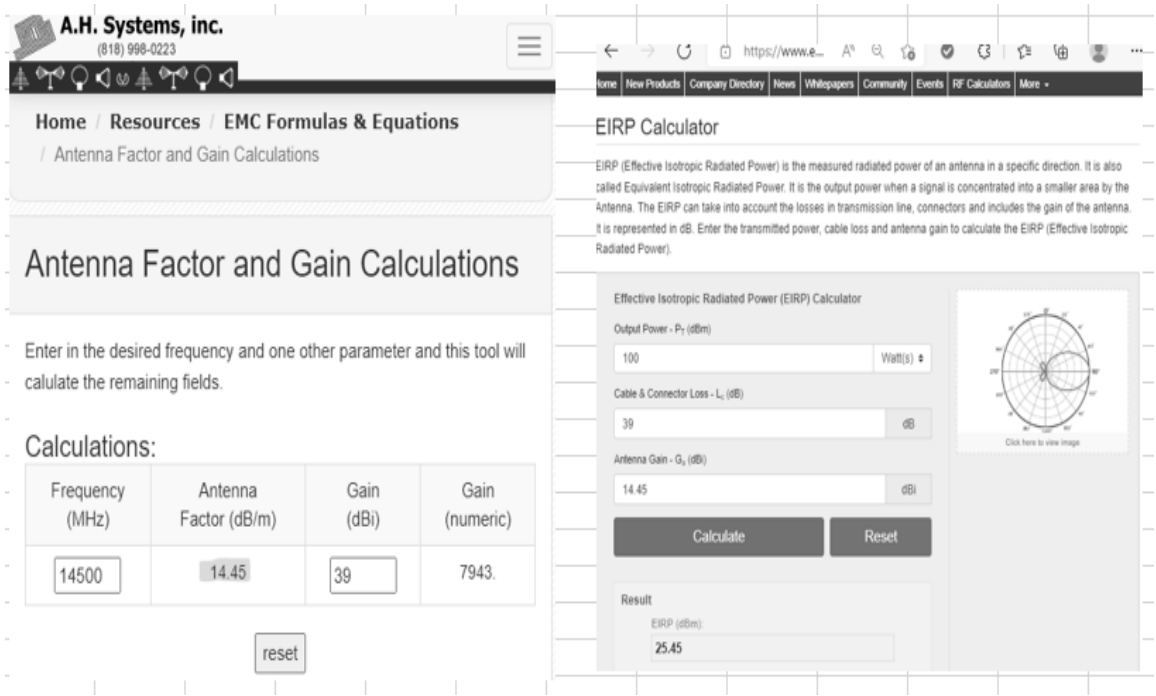


Figure 3: Calculations for Starwin High Gain Portable

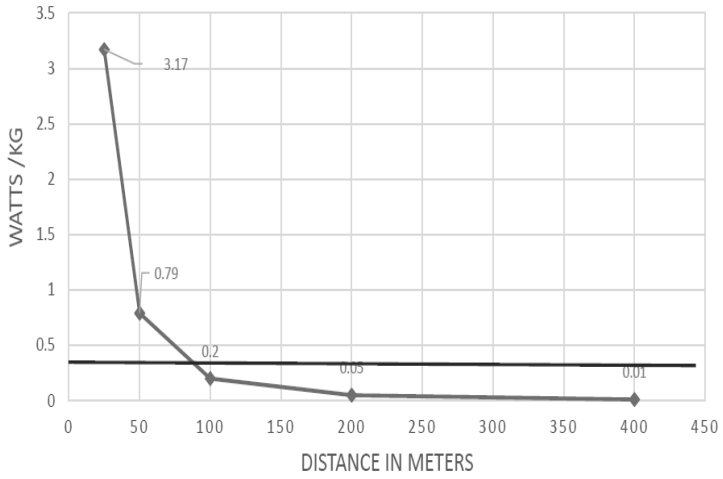


Figure 4: GX 60 SAR W/KG

Hypothesis $H_{a,\text{ the value of SAR,}}$ is met at a distance of fewer than 100 meters. This value provides validation while showing that a cruise ship could become a significant area target where people can be radiated.

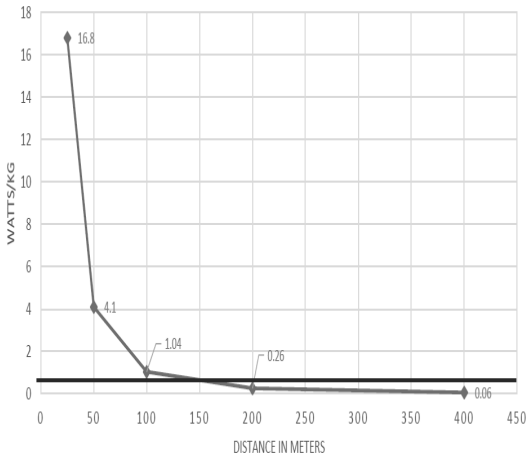


Figure 5: GX 100 SAR W/KG

Hypothesis H_a, the value of SAR, is met at a distance of fewer than 160 meters. This value provides validation while showing that a cruise ship could become a prominent area target where people can be radiated.

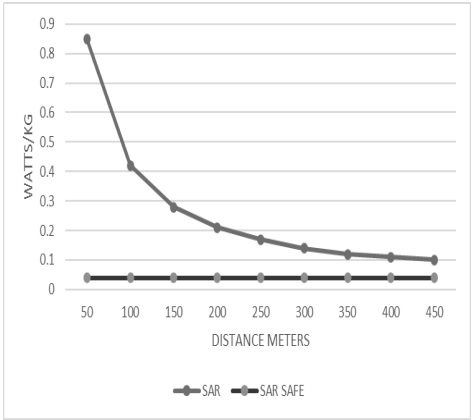


Figure 6: L3Harris HackeyeTM 4 Lite

Hypothesis H_a, the value of SAR, is met at a distance of fewer than 400 meters. This value provides validation while showing that a land-based system could be used to target specific areas.

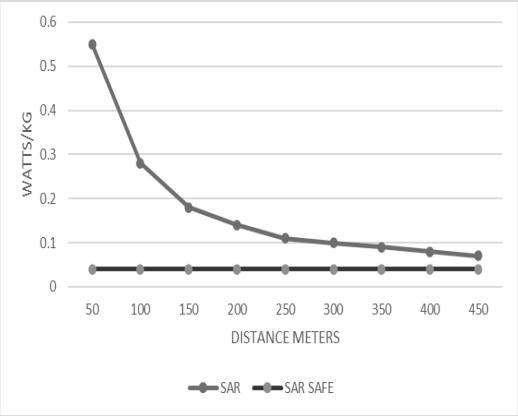


Figure 7: Starwin High Gain Terminal

Hypothesis H_a, the value of SAR, is met at a distance of fewer than 150 meters. This value provides validation while showing that a land-based system could be used to target specific areas.

The results of this replication study corroborate what the original author provided as a hypothesis for the dangers of specific absorption rates (SAR). All the systems studied in this virtual environment provided feedback of power that shows serious harm can come to biological tissue; in most cases, human beings, due to the high SAR, at times, the values were hundreds of times higher than the safe limit of exposure. Further adds to the theory that satellite communication equipment can be turned into a weapon given enough knowledge of radio wave theory, cybersecurity, the location of the attack, and the intended target. Static locations such as ships, buildings, or open recreational areas can be flooded with radio wave radiation that can, over time, cause illness to those that are continuously exposed. One example of this attack was conducted at the United States Embassy in Habana, Cuba. As per the image below, the distance from the embassy to other buildings is less than 38 meters, well within the highest output radius of all tested systems. If an antenna is pointed at the facility for a prolonged period, personnel assigned to that embassy can start showing symptoms related to radio wave radiation.

United States Embassy, La Habana Cuba



Figure 8: map calculator, 2022



Figure 9: google maps, 2022

8. GAP / Limitations

There are some limitations to this replication study. While our replication study attempted to address the gaps and build on the authors' recommendations McKay (2021), this study had notable limitations. First, just like in the author's original work, we could not conduct a physical test of the SATCOM equipment. Studies like this should be extended from simulation to field testing. Second, due to the software's proprietary nature, we cannot thoroughly test the hardening of the software of each system. Third, manufacturers did not provide the maximum tolerance for the high-power amplifiers, thus making it unknown if the equipment could withstand an overload before becoming inoperable. In these scenarios, most attackers do not concern themselves with how

long equipment may operate. However, when long-term overt and denial actions are taken, the need to ensure the viability of higher-end tolerance is required.

9. Conclusion

In conclusion, our primary goal in conducting this replication study was to evaluate the robustness of cyber-physical attacks using high-intensity radiated fields. Our replication shows that the results of McKay (2021) study are exciting and helpful for future research. After completing our study, we feel that it is evident that there is much scope for research on cyber-physical attacks. It is paramount to be aware that the satellite system's cybersecurity challenges are enormous in cybersecurity and human defense. Furthermore, the maritime or aerospace industry has also developed a lot so have the risks that weigh on them. This replication study reinforces the argument that further studies are needed on this matter and greater cooperation between manufacturers and cybersecurity analysts.

Definition of Terms

- Decibel (dB). Ten times the logarithm to the base ten of the ratio of two power levels (OET 65 (1997), pp.2-5).
- EIRP (Effective Isotropic Radiated Power) is the measured radiated power of an antenna in a specific direction
- Gain (of an antenna) is a ratio that is usually expressed in decibels of the power required at the input of a loss-free reference antenna to the power supplied to the input of the given antenna to produce, in a given direction, the same field strength or the same power density at the same distance. When not specified otherwise, the gain refers to the direction of maximum radiation. Gain may be considered for a specified polarization. Gain may be referenced to an isotropic antenna (dBi) or a half-wave dipole (dBd) (McKay (2021), pp 44; OET 65 (1997), pp.2-5).
- Hertz (Hz). The unit for expressing frequency, (f). One hertz equals one cycle per second. Magnetic field strength (H). A field vector equals the magnetic flux density divided by the medium's permeability (OET 65 (1997), pp.2-5).
- High-Intensity Radiated Fields (HIRF) are those that can produce, within the frequency domain from 10kHz to 40 GHz, an electromagnetic field strength sufficient to cause a SAR rating of 0.4 w/kg or above (McKay, 2021).
- SATCOM: a global Wi-Fi network, the roots of various services used on vessels to guarantee safety and security (Caprolu et al., 2020).
- Specific absorption rate (SAR) measures the energy absorbed by (dissipated in) a total mass contained in a volume element of dielectric materials such as biological tissues. It is usually expressed in watts per kilogram (W/kg) or per gram (mW/g). Human exposure to RF fields is based on SAR thresholds where adverse biological effects may occur. When the human body is exposed to an RF field, the SAR experienced is proportional to the squared value of the electric field strength induced in the body (McKay (2021) , pp 48; OET 65 (1997), pp.2-5).

References

Arampatzis, A., 2021. U.S. national cybersecurity plan to SAFEGUARD Maritime Sector https://www.tripwire.com/state-of-security/security-data-protection/cyber-security/us-national-cybersecurity-plan-safeguard-maritime-sector/

Baumann, S., et al., 1997. The electrical conductivity of human cerebrospinal fluid at body temperature. IEEE Transactions on Biomedical Engineering, 44(3), 220-223. doi:10.1109/10.554770

Caprolu, M., et al., 2020. Vessels Cybersecurity: Issues, Challenges, and the Road Ahead. IEEE Communications Magazine, 58(6), 90-96. doi:10.1109/mcom.001.1900632

Cleveland, R. F., Jr., 2001. Evaluating Compliance with FCC Guidelines for Human Exposure to Radiofrequency Electromagnetic Fieldshttps://transition.fcc.gov/Bureaus/Engineering_Technology/Documents/bulletins/oet65/oet65.pdf

Electromagnetic Fields (EMF)., 2016. Retrieved from https://www.icnirp.org/cms/upload/publications/ICNIRPrfgdl2020.pdf

FCC. Radio Frequency Safety., 2017. Retrieved from https://www.fcc.gov/general/radio-frequency-safety-0#block-menu-block-4'

Giri, D., 2004. High-Power Electromagnetic Radiators: Nonlethal Weapons and Other Applications. Harvard University Press. Cambridge, Massachusetts. 2004

Greenburg, M. and Chalk, P., 2006. Maritime Terrorism: Risk and Liability. RAND Center for Terrorism and Risk Management Policy. Arlington, Virginia

Gurevich, V., 2005. Electromagnetic Terrorism: New Hazards. Israel Electric corp., Central Electric Laboratory. Haifa, Isreal.http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/11280/1/EE_2005_4_Gurevich_%D0%95lectromagnetic.pdf

Hudaib, A. A., 2016. Satellite Network Hacking & Security Analysis. International Journal of Computer Science and Security, 10(1), 8-55.

Livingston, D. and Lewis P., 2016. Space the Final Frontier for Cybersecurity. Chatham House. Retrieved from https://www.chathamhouse.org/sites/default/files/publications/research/2016-09-22- space-final-frontier-cybersecurity-Livingstone-lewis.pdf

Loukas, G., 2015. Cyber-Physical Attacks: A Growing Invisible Threat. Butterworth- Heinemann. Oxford, United Kingdom

Marlatt, G., 2005. Directed Energy Weapons (DEWs): A Bibliography. Calhoun NPS.https://calhoun.nps.edu/bitstream/handle/10945/6980/Oct05- DEW_biblio.pdf?sequence=1

Manz, B., 2017. RF Energy Is Finally Cooking. Retrieved from https://www.mwrf.com/community/article/21848854/rf-energy-is-finally- cooking https://www.mwrf.com/community/article/21848854/rf-energy-is-finally-cooking

McKay, R. W., 2021. Cyber-Physical Attack Using High-Intensity Radiated Fields (Doctoral dissertation, Marymount University).

National Academies of Sciences, Engineering, and Medicine., 2020. An assessment of illness in U.S. government employees and their families at overseas embassies. Washington, DC: The National Academies Press. https://doi.org/10.17226/25889

National Post Staff., 2020. China used 'microwave' pulse weapon to Force Indian soldiers Off HIMALAYAN hilltops: Report. Retrieved from https://www.msn.com/en-ca/news/world/china-used-microwave-pulse-weapon-to-force- Indian-soldiers-off-Himalayan-hilltops-report/ar-BB1b5QtW

OET 56, Federal Communications Commission., 1999. Questions and Answers about Biological Effects and ... Retrieved from https://transition.fcc.gov/Bureaus/Engineering_Technology/Documents/bulletins/oet56/o et56e4.pdf

OET 65, Federal Communications Commission., 1997. Evaluating Compliance with FCC Guidelines for Human Exposure to Radiofrequency Electromagnetic Fields. Retrieved from https://transition.fcc.gov/bureaus/oet/info/documents/bulletins/oet65/oet65.pdf

Rooker, JW., 2008. Satellite Vulnerabilities. Defense Technical Information Center. Retrieved from http://www.dtic.mil/docs/citations/ADA507952

Santamarta, R., 2018. Last Call for SATCOM Security - ioactive.com. Retrieved from https://ioactive.com/wp-content/uploads/2018/08/us-18- Santamarta-Last-Call-For-Satcom-Security-wp.pdf

Santamarta, R., 2014. A Wake-up call for SATCOM Security – ioactive.com. Retrieved from https://ioactive.com/pdfs/IOActive_SATCOM_Security_WhitePaper.pdf

Santamarta, R., 2014. SATCOM Terminals Hacking by Air, Sea, and Land –ioactive.com. Retrieved from http://blackhat.com/docs/us-14/materials/us-14-Santamarta-SATCOM-Terminals-Hacking-By-Air-Sea-And-Land-WP.pdf

United States Navy. Radio frequency-radiation. Retrieved from https://www.public.navy.mil/navsafecen/pages/acquisition/radio_frequency- radiation.aspx

White House Trump Administration., 2020. National archives and Records Administration. Retrieved from https://trumpwhitehouse.archives.gov/briefings-statements/statement-national-security- advisor-Robert-c-Obrien-regarding-national-maritime-cybersecurity-plan/

National Academy of Sciences and Medicine., 2020. National Academy of Sciences and Medicine', Retrieved fromFront Matter | An Assessment of Illness in U.S. Government Employees and Their Families at Overseas Embassies |The National Academies Press

Cybersecurity in Digital Transformation Applications: Analysis of Past Research and Future Directions

Zakariya Belkhamza
Ahmed Bin Mohammed Military College, Doha, Qatar

zbelkhamza@abmmc.edu.qa

Abstract: The term *digital* is often used to indicate the changes occurring in today's world, generally referred to as a cyber-physical system, driven by the rapid adoption of digital technologies, where the cyber and the physical worlds are partly overlapping. Digital transformation refers to the integration of the digital technology of the cyber world into all physical domains. These cyber-physical systems must be secure against the threat of cyberattacks. However, one of the most challenging aspects of cybersecurity is the evolving nature of cyberattack risks, which is highly integrated with digital transformation. Despite a number of published articles, there is little investigation of past literature analysis that presents digital transformation applications and cybersecurity trends. The objective of this paper is to provide an intensive examination of digital transformation applications and cybersecurity research between 2019 and 2022, to detect the most profound research areas, emphasize existing challenges and identify patterns, tendencies or regularities existing in the literature in terms of technological applications. This aims to support scholars with a comprehensive understanding of the past, present and future directions of this research trend. To achieve these objectives, a systematic literature review is utilized. The findings introduce several implications for the present state of the literature, apparent study gaps and several research questions, which can be explored in future research.

Keywords: digital transformation, cybersecurity, systematic literature review, systematic analysis

1. Introduction

Cybersecurity has become a global challenge for digital business activities. Research on cybersecurity in the digital transformation framework has gained significant attention in recent years (Mendhurwar & Mishra, 2021; Sandhu, 2021; Swain et al., 2022). Some research has paid specific attention to the subject of cybersecurity in the post COVID-19 era. For example, Alawida et al. (2022) reviewed and examined the effect of cybersecurity on organizations during the global COVID-19 crisis. Others have focused on various areas of digital transformation, such as cybersecurity in blockchain-based systems (Schlatt et al., 2022), big data analytics (Rawat, Doku & Garuba, 2021), machine learning (Aiyanyo, Samuel & Lim, 2020) and the Internet of things (Rudenko et al., 2022).

Despite these investigations and reviews, and a few recent studies that conducted a literature analysis (Carcary, Doherty & Conway, 2019; Zhu, Ge, & Wang, 2021), a comprehensive analysis of cybersecurity in the digital transformation context still lacks further investigation. Therefore, there is an apparent need for research that provides a comprehensive understanding of the past, present and future of research on cybersecurity in the digital transformation context.

This paper addresses three research questions: 1) what is the current status of cybersecurity research? 2) what are the key areas that interlink cybersecurity and digital transformation applications? and 3) what potential future research avenues can be identified based on existing literature that concern applied technologies of digital transformation? To provide answers to these questions, a systematic review is presented.

2. Research Methodology

To answer the research questions set for this study, a systematic literature review was conducted. Although there are several guidelines for conducting a literature review, this study followed the guidelines published by Kitchenham and Charters (2007). Several studies have successfully used this guideline, for example, Taylor et al. (2020). The following section reports the methodology of this study.

2.1 Inclusion criteria

Since the research article is the unit of analysis, it is important before engaging in a study to set the article inclusion criteria. We used Scopus database as the platform to collect the sample of this research. First, although it was not necessary for the study to directly address the post COVID-19 period, articles published between 2019–2022 were targeted for this research. Second, only articles published in English were considered. Third, articles published in irrelevant research areas, such as law and medicine, were excluded. This refinement process resulted in a total of 67 articles in the Scopus database; after filtering the sample for the database, duplicate articles were excluded, and a final sample of 65 articles was selected.

2.2 Quality assessment

To maintain a quality review for this study, it was important to assess the quality of primary studies as recommended by Kitchenham and Charters (2007). The objective of this stage was to assess the relevance of the papers with regard to the research questions. The following rules were set and each article from the sample had to fulfil these to be included in the study:

Stage 1: **Cybersecurity.** The article had to focus on the use of cybersecurity or the application of cybersecurity technology to a specific problem and be well-commented.

Stage 2: **Digital Transformation.** The article had to focus on digital technology or the various types of digital transformation.

Stage 3: The article had to address the issue of cybersecurity in direct accordance with digital transformation.

Stage 4: The article had to provide sufficient context for the research objectives and findings. This would allow for an accurate interpretation of the research.

Stage 5: The article had to provide sufficient context in the study to enable the findings to be generalized to a larger context in the research field.

Stage 6: Details of how the study was conducted and performed had to be explained and reported in order to determine accuracy.

Each article was carefully examined against the above checklist, and articles that failed to meet all the above criteria were removed from the sample. It was found that eight studies did not meet one or more of the checklist items and, therefore, were removed from the sample. Two articles were found to be duplicated, as their findings were published in another article; therefore, these two duplicated articles were excluded. The final sample of the study consisted of 18 papers that passed the quality assessment. Information regarding the articles included in the research is summarized in Table 1.

Table 1: Summary of research articles

Study	Problems of study	Nature of study	Objectives of study	Implications of study
Castelo-Branco et al. (2022)	Inconsistency of available maturity models beyond the manufacturing industry.	Mixed-method approach.	Presented an Industry 4.0 maturity index.	Industry 4.0 implementation to be integrated into a consistent strategic decision framework that went beyond technology adoption and was market oriented.
Haleem et al. (2022)		Literature review analysis.	Identified the major cybersecurity applications for Industry 4.0 and discussed them.	Digital transformation in the industry 4.0 era would increase industrial competitiveness and improve their capacity to make optimum decisions.
Rodrigues et al. (2022)	Issue of artificial intelligence (AI), digital transformation and cybersecurity in the finance and banking sector.	Mixed-method approach.	Developed a realistic decision-support model by combining cognitive mapping and the decision-making trial and evaluation laboratory (DEMATEL) method.	Artificial intelligence, digital transformation and cybersecurity would allow for a better decision-making process in terms of security, efficiency and efficacy in ICT ecosystems.
Malatji, Marnewick and Von Solms (2022)	Interconnectivity between enterprise information technology (IT) and industrial control systems (ICS) environment.	Literature review analysis.	Developed a critical infrastructure cybersecurity capability framework.	Developed a critical infrastructure cybersecurity capability framework comprising 29 capability domains (cybersecurity focus areas).
Reina Quintero et al. (2022)	Existing access control models failed to integrate sophisticated decision-making in security policies.	Design analysis.	Presented a model-based approach to facilitate the definition of complex security policies based on usage control (UCON) in the software development process.	Improved security policies beyond simple access control rules and covered more complex and dynamic scenarios and provided powered decision-making ability.
Muncinelli et al. (2021)	Issue of digital compliance with legal requirements, risks,	Exploratory analysis.	Analysed the primary areas of contribution to the assessment of	Presented the components of the MCP-LGPD capability model (process capability

Study	Problems of study	Nature of study	Objectives of study	Implications of study
	business analysis, good practices and standards efficiently and sustainably.		process capability for digital transformation concerning cybersecurity in the context of personal data protection legislation.	model-Brazil's General Data Protection Law).
Sonkor and De Soto (2021)		Literature review analysis.	Understood the current state of the art and identified gaps to suggest future directions regarding operational technology in construction from the perspective of cybersecurity.	Investigated methods to evaluate security levels on construction sites and countermeasures against cyberattacks.
Petratos (2021)	Insufficient examination of the consequences of misleading information on businesses.	Exploratory analysis.	Offered a primer on misleading information and cyber risks aimed at business executives and leaders across an array of industries, organizations and nations.	Established cybersecurity policies and best practices to manage emerging and multifaceted cyberthreats.
Tarikere, Donner and Woods (2021)	Critical need for cybersecurity in the development of Internet of Medical Things (IoMT) technology, alongside design and utility.	Exploratory analysis.	Examined the market opportunities and risks associated with IoMT.	Implementation of risk management plans would ensure continuity of care that embraced technologies while protecting patient data and information.
Lee (2021)	Rising issues of cyber breaches and the emergence of novel cybersecurity technologies such as machine learning and AI.	Exploratory analysis.	Introduced a cyber risk management framework and illustrated a continuous improvement of cybersecurity performance and cyber investment cost analysis.	Developed a cyber risk management framework in which risk management activities were organized and evaluated in four layers.
Mironeanu et al. (2021)	Increasing number of threats and attack patterns in real-time analytical tools.	Design analysis.	Introduced a novel concept for integrating machine learning and analytical tools into a live intrusion detection and prevention solution.	Proposed the Experimental Cyber Attack Detection Framework (ECAD).
Emer, Unterhofer and Rauch (2021)	Issue of digitalization and connectivity in a digital environment where human–machine collaboration is more flexible and agile.	Exploratory analysis.	Introduced an assessment model of Small & Medium Enterprise cybersecurity for better managerial action and data security.	Concept of an easy-to-use cybersecurity assessment tool would strongly improve the currently difficult introduction of cyber protection solutions in SMEs.
Villegas et al. (2021)		Literature review analysis.	Conducted a literature review of technological pillars to implement Industry 4.0.	Proposed strategies to successfully integrate AI and the technological pillars of Industry 4.0.
Yerina et al. (2021)	Monitoring cyber incidents and anomalies for cybersecurity assessment.	Empirical analysis.	Summarized the international experience of assessing cybersecurity and the effective protection of cyberspace at a national level.	Cybersecurity ratings played the role, to some extent, of an identifier of the relative advantages and vulnerabilities of national cyber strategies and indicated the need for their review to strengthen protection against cyber-attacks.
Mendhurwar and Mishra (2021)	Integration of modern technologies facing key challenges related to information security and privacy.	Exploratory analysis.	Explored interplay and synergetic relationships among technologies and identified key challenges on cybersecurity and privacy.	Identified opportunities related to the integration of social and IoT technologies.
Uddin et al. (2020)	The more banks improve their operational efficiency and quality of service by relying on	Empirical analysis.	Examined whether the law of diminishing marginal returns from overspending on cyber	Established aggressive cyber technology spending could lead to diminishing returns,

Study	Problems of study	Nature of study	Objectives of study	Implications of study
	cyber technology the more they become vulnerable to cybersecurity.		technology affected bank stability.	which adversely affected stability of banks.
Chatfield and Reddick (2019)	Strategic use of the IoT is still in the early stages of development in governments across the globe (Meyers et al., 2015) and the IoT adoption rate remains low.	Literature review analysis.	Developed a new framework for IoT-enabled smart government.	Addressed knowledge gap in the literature on smart government, from a dynamic capabilities theory perspective.
Yuan et al. (2022)	Impact of various power application services on the security situation still requires further investigation.	Design analysis.	Proposed a method for quantifying the information security of the power network based on the evolutionary neural network.	Improved the quantification accuracy of the information security of the power network to a certain extent.

3. Results and Discussion

It is significant to observe how digital transformation is introduced and presented in the literature. Most studies failed to consider defining the operationalization of digital transformation in the context of the study. Various articles presented digital transformation as a process, which is a synonym of the so-called Industry 4.0 (Castelo-Branco et al., 2022). Other studies investigated elements of digital transformation, such as IoT (Chatfield & Reddick, 2019; Mendhurwar & Mishra, 2021; Villegas et al., 2021), cyber-physical systems (Sonkor & De Soto, 2021) and risk management (Lee, 2021; Uddin, Mollah, & Ali, 2020).

Rodrigues et al. (2022), for instance, differentiated between artificial intelligence and digital transformation, and adopted Kaplan & Haenlein (2019)'s definition as *'a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation'*. However, they defined digital transformation simply as the use of digital technology in new ways to solve traditional problems.

Unlike digital transformation, cybersecurity is well-defined in the literature; for instance, Lee (2021) defined cybersecurity as the preservation of the confidentiality, integrity and availability of information in complex environments resulting from the interaction of people, software and services on the Internet by means of technology devices and networks connected to it. Several studies investigated components of cybersecurity such as cyber threat, cyberattack, cybercrime and cyber risk.

Table 2: Types of methodology of articles

Methodology of articles	Number of articles	%
Exploratory analysis	6	34%
Literature review analysis	5	28%
Design analysis	3	16%
Mixed-method analysis	2	11%
Exploratory analysis	2	11%
Total	18	100%

From a methodological perspective, the articles may be broadly grouped into five categories: the first group used exploratory analysis and consisted of six articles; the second group used literature review analysis and consisted of five articles; the third group used design analysis and consisted of three articles; and two groups consisted of two articles each, and used mixed-method analysis and empirical analysis, respectively. Table 2 summarizes these findings.

In terms of the nature of research, the sample can be grouped into two groups equally with nine articles each; the first group of articles proposed model/framework development, and the second represented conceptual investigation.

Further insight into the first group of articles revealed that five articles developed some interesting frameworks that could be used in different scenarios in cybersecurity. These frameworks are: the Industry 4.0 maturity framework (Castelo-Branco et al., 2022), the critical infrastructure cybersecurity capability framework (Malatji, Marnewick & Von Solms, 2022), the cyber risk management framework (Lee, 2021), the experimental

cyberattack detection framework (Mironeanu et al., 2021) and the IoT enabled smart government framework (Chatfield & Reddick, 2019). In addition to these frameworks, two models were proposed: the realistic-decision support model (Rodrigues et al., 2022) and the SME cybersecurity assessment model (Emer, Unterhofer & Rauch, 2021), and a model for quantifying the information security of the power network (Yuan et al., 2022).

Finally, these studies suggested that the most technological applications were IoT (Chatfield & Reddick, 2019; Mendhurwar & Mishra, 2021; Tarikere, Donner & Woods, 2021; Villegas et al., 2021), AI (Mendhurwar & Mishra, 2021; Rodrigues et al., 2022; Villegas et al., 2021), robotics and virtual reality (Villegas et al., 2021).

4. Future Research Directions

This article reports on the cybersecurity in digital transformation applications. Based on the results of this study, some research directions concerning cybersecurity are presented. Cybersecurity is well positioned in the literature, and various aspects of it have been investigated in various applications of technology. However, a clear framework of cybersecurity in the digital transformation process is still required, and most of the literature investigated few aspects of digital transformation, while an inclusive framework of digital transformation was not observed. Thus, future research needs to develop comprehensive and quantifiable guidelines to fill this gap in the literature. Moreover, the findings from this study are subject to the inherent limitations of a small sample. The sample of the study was limited to journal articles in English from the Scopus database, and this might not provide an adequate generalizability of the findings. Future studies may address these limitations by exploring different databases, including book chapters and conference proceedings, and may also include documents in other languages. Finally, this study only focused on articles published between 2019–2022, a period that might not actually represent the current research in the domain. Further studies may expand the period included to afford a better insight. From this systematic review, it is concluded that a collaborative and standardized approach for the application of various digital transformation manifestations on cybersecurity attacks is required to collectively strengthen organizations against increasing cyber threats.

References

Aiyanyo, I. D., Samuel, H., and Lim, H. (2020)."A Systematic Review of Defensive and Offensive Cybersecurity with Machine Learning", *Applied Sciences (Switzerland)*, Vol. 10, No. 17. https://doi.org/10.3390/app10175811

Alawida, M., Omolara, A. E., Abiodun, O. I., and Al-Rajab, M. (2022)."A Deeper Look into Cybersecurity Issues in the Wake of Covid-19: A Survey", *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.08.003

Carcary, M., Doherty, E., and Conway, G. (2019)."A Framework for Managing Cybersecurity Effectiveness in the Digital Context", *European Conference on Information Warfare and Security, ECCWS*, *2019-July*, pp 78–86.

Castelo-Branco, I., Oliveira, T., Simões-Coelho, P., Portugal, J., and Filipe, I. (2022)."Measuring the Fourth Industrial Revolution Through the Industry 4.0 Lens: The Relevance of Resources, Capabilities and the Value Chain", *Computers in Industry*, Vol. 138. https://doi.org/10.1016/j.compind.2022.103639

Chatfield, A.T., and Reddick, C. G. (2019)."A Framework for Internet of Things-Enabled Smart Government: A Case of IoT Cybersecurity Policies and Use Cases in U.S. Federal Government", *Government Information Quarterly*, Vol. 36, No. 2, pp 346–357. https://doi.org/10.1016/j.giq.2018.09.007

Emer, A., Unterhofer, M., and Rauch, E. (2021)."A Cybersecurity Assessment Model for Small and Medium-Sized Enterprises", *IEEE Engineering Management Review*, Vol. 49, No. 2, pp 98–109. https://doi.org/10.1109/EMR.2021.3078077

Haleem, A., Javaid, M., Singh, R. P., Rab, S., and Suman, R. (2022)."Perspectives of Cybersecurity for Ameliorative Industry 4.0 Era: A Review-Based Framework", *Industrial Robot*, Vol. 49, No. 3, pp 582–597. https://doi.org/10.1108/IR-10-2021-0243

Kaplan, A., and Haenlein, M. (2019)."Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence", *Business Horizons*, Vol. 62, No. 1, pp 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kitchenham, B. and Charters, S. (2007). Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report

Lee, I. (2021)."Cybersecurity: Risk Management Framework and Investment Cost Analysis", *Business Horizons*, Vol. 64, No. 5, pp 659–671. https://doi.org/10.1016/j.bushor.2021.02.022

Malatji, M., Marnewick, A. L., and Von Solms, S. (2022)."Cybersecurity Capabilities for Critical Infrastructure Resilience", *Information and Computer Security*, Vol. 30, No. 2, pp 255–279. https://doi.org/10.1108/ICS-06-2021-0091

Mendhurwar, S., and Mishra, R. (2021)."Integration of Social and IoT Technologies: Architectural Framework for Digital Transformation and Cyber Security Challenges", *Enterprise Information Systems*, Vol. 15, No. 4, pp 565–584. https://doi.org/10.1080/17517575.2019.1600041

Mironeanu, C., Archip, A., Amarandei, C.-M., and Craus, M. (2021)."Experimental Cyber Attack Detection Framework. *Electronics (Switzerland)*, Vol. 10, No. 14. https://doi.org/10.3390/electronics10141682

Muncinelli, G., De Lima, E. P., Cestari, J. M. A. P., Deschamps, F., and Da Costa, S. E. G. (2021)."Developing a conceptual model for process capability in the brazilian data protection regulation context". *Journal of Industrial Integration and Management*, Vol. 6, No. 4, pp 407–427. https://doi.org/10.1142/S2424862221400017

Petratos, P. N. (2021)."Misinformation, disinformation, and fake news: Cyber risks to business". *Business Horizons*, Vol. 64, No. 6, pp 763–774. https://doi.org/10.1016/j.bushor.2021.07.012

Rawat, D. B., Doku, R., and Garuba, M. (2021)."Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security", *IEEE Transactions on Services Computing*, Vol. 14, No. 6, pp 2055–2072. https://doi.org/10.1109/TSC.2019.2907247

Reina Quintero, A. M., Pérez, S. M., Varela-Vaca, Á. J., López, M. T. G., and Cabot, J. (2022)."A domain-specific language for the specification of UCON policies". *Journal of Information Security and Applications*, Vol. *64*. https://doi.org/10.1016/j.jisa.2021.103006

Rodrigues, A. R. D., Ferreira, F. A. F., Teixeira, F. J. C. S. N., and Zopounidis, C. (2022). "Artificial Intelligence, Digital Transformation and Cybersecurity in the Banking Sector: A Multi-Stakeholder Cognition-Driven Framework", *Research in International Business and Finance*, Vol. 60. https://doi.org/10.1016/j.ribaf.2022.101616

Rudenko, R., Pires, I. M., Oliveira, P., Barroso, J., and Reis, A. (2022)."A Brief Review on Internet of Things, Industry 4.0 and Cybersecurity", *Electronics (Switzerland)*, Vol. 11, No. 11. https://doi.org/10.3390/electronics11111742

Sandhu, K. (2021)."Advancing Cybersecurity for Digital Transformation: Opportunities and Challenges", *Handbook of Research on Advancing Cybersecurity for Digital Transformation*, pp 1–17. https://doi.org/10.4018/978-1-7998-6975-7.ch001

Schlatt, V., Guggenberger, T., Schmid, J., and Urbach, N. (2022)."Attacking the Trust Machine: Developing an Information Systems Research Agenda for Blockchain Cybersecurity", *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2022.102470

Sonkor, M. S., and García De Soto, B. (2021)."Operational Technology on Construction Sites: A Review from the Cybersecurity Perspective", *Journal of Construction Engineering and Management*, Vol. 147, No. 12. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002193

Swain, A., Swain, K. P., Pattnaik, S. K., Samal, S. R., and Das, J. K. (2022)."Cybersecurity in Digital Transformations", *Lecture Notes in Networks and Systems*, Vol. 430, 247–252. https://doi.org/10.1007/978-981-19-0825-5_26

Tarikere, S., Donner, I., and Woods, D. (2021)."Diagnosing a Healthcare Cybersecurity Crisis: The Impact of IoMT Advancements and 5G", *Business Horizons*, Vol. 64, No. 6, pp 799–807. https://doi.org/10.1016/j.bushor.2021.07.015

Taylor, P. J., Dargahi, T., Dehghantanha, A., Parizi, R. M., and Choo, K. K. R. (2020)."A systematic literature review of blockchain cyber security". *Digital Communications and Networks*, Vol. 6, No. 2, pp 147–156. https://doi.org/10.1016/j.dcan.2019.01.005

Uddin, M. H. M. H., Mollah, S., and Ali, M. H. M. H. (2020). "Does Cyber Tech Spending Matter for Bank Stability?" *International Review of Financial Analysis*, Vol. 72. https://doi.org/10.1016/j.irfa.2020.101587

Villegas, O. O. V., Nandayapa, M., Azuela, J. H. S., Franco, E. G. C., and Linares, G. T. R. (2021). ”Artificial Intelligence for Industry 4.0 in Iberoamerica", *Computacion y Sistemas*, Vol. 25, No. 4, pp 761–773. https://doi.org/10.13053/CyS-25-4-4056

Yerina, A. M., Honchar, I. A., and Zaiets, S. V. (2021)."Statistical indicators of cybersecurity development in the context of digital transformation of economy and society". *Science and Innovation*, Vol. *17*, No. 3, pp 3–13. https://doi.org/10.15407/scine17.03.003

Yuan, Q., Pi, Y., Kou, L., Zhang, F., and Ye, B. (2022)."Quantitative Method for Security Situation of the Power Information Network Based on the Evolutionary Neural Network. *Frontiers in Energy Research*, Vol. 10. https://doi.org/10.3389/fenrg.2022.885351

Zhu, X., Ge, S., and Wang, N. (2021)."Digital Transformation: A Systematic Literature Review", *Computers and Industrial Engineering*, Vol. 162. https://doi.org/10.1016/j.cie.2021.107774

An Automated Post-Exploitation Model for Cyber Red Teaming

Ryan Benito, Alan Shaffer and Gurminder Singh
Naval Postgraduate School, Monterey, California, USA

ryan.m.benito.mil@us.navy.mil
alan.shaffer@nps.edu
gsingh@nps.edu

Abstract: Red teaming is a well-established methodology for ensuring and augmenting cyber system security; however, the training, expertise, and knowledge of appropriate tools and techniques required to perform effective red teaming come with a significant cost in time and resources. Large organizations such as the Department of Defense (DOD) use vulnerability assessment to identify software patches and other remediations for cyber systems to mitigate cyberspace exploitation. If a patch cannot be applied in a timely manner, for instance to minimize network downtime, measuring and identifying the impact of such unpatched vulnerabilities is left to scarce red teaming services. These services typically concentrate on initial access exploitation, which stops short of exploring the larger security impacts of cyber threats performing post-exploitation actions. This gap in post-exploitation red team analysis results in increased susceptibility to adversary offensive cyberspace operations (OCO) against DOD systems. This research extends the Cyber Automated Red Team Tool (CARTT), developed at the Naval Postgraduate School, by implementing automated red team post-exploitation analysis. The intent of this extended capability is to reduce the workload on limited DOD red teams and penetration testers by providing system administrators with the ability to perform deeper system analysis for the impacts of exploited vulnerabilities.

Keywords: Red teaming, automated cyber post-exploitation, defensive cyber operations, web-based assessment.

1. Introduction

The FY23 U.S. Department of Defense (DOD) budget increased funding for "cyberspace efforts" by 8% to $11.2 billion, indicating a greater need for cyberspace operations (CO) resources. Although this is a significant increase, CO resources like cyber protection teams (CPT), cyber mission teams (CMT), red teams, and national mission teams (NMT), all of which conduct offensive cyberspace operations (OCO), defensive cyberspace operations (DCO), and DOD information network (DODIN) operations, continue to be scarce throughout the DOD. Moreover, the DOD requires significant investment in time, skill, resources, and human capital to provide critical cyberspace insights, such as vulnerability assessments, threat emulation, and network hardening.

The scarcity of CO resources may leave low priority organizations without critical cybersecurity support, forcing them to wait for resources, and to potentially remain vulnerable to cyberspace attack or exploitation. DOD organizations do not currently have a tool that enables non-experts to self-assess their systems to triage and prioritize which vulnerabilities should be patched based on relative risk.

Prior to the work reported here, previous research and development on the Cyber Automated Red Team Tool (CARTT) at the Naval Postgraduate School in Monterey, CA, has automated the process for vulnerability assessment and some aspects of initial access exploitation, enabling users to assess the success or failure of a given initial access exploit against hosts on their systems. However, CARTT did not provide the ability to conduct post-exploitation assessment of these systems (Berrios 2020).

This paper summarizes research to develop and implement an automated post-exploitation capability for CARTT. We also summarize an impact-based post-exploitation taxonomy derived from the MITRE ATT&CK® framework to help organizations more accurately identify vulnerability impact and risk in implementing security controls (Benito 2022).

The rest of this paper is organized into five sections: related work, an impact-based post-exploitation taxonomy, system design, system implementation, and a summary of conclusions and future work.

2. Background

2.1 Cyber Post-Exploitation

Post-exploitation actions occur after a cyberspace attacker gains initial access onto an adversary's system. Without initial access exploitation, post-exploitation actions would not be feasible. Firewalls and intrusion detection and prevention systems can support perimeter defense strategies by keeping adversaries out of the network, however, a major vulnerability of a perimeter defense strategy is assuming that users within the

perimeter can always be trusted. Adversaries that can pierce perimeter defenses may exploit this assumed trust to perform post-exploitation actions.

The National Institute of Standards and Technology (NIST) and the DOD have created the Zero Trust Architecture (ZTA) and framework to thwart the implicit trust granted to assets or user accounts based solely on their physical or network location or based on asset ownership (DOD 2021; Rose et al 2020). Zero trust is not definitive though. Motivated adversaries will continue to find creative ways to conduct post-exploitation despite new security measures offered by a ZTA implementation. The tools, techniques, and procedures of post-exploitation actions can provide the necessary indicators of compromise to enable operators to identify and eradicate adversaries from the network. Some post-exploitation actions can include persistence, privilege escalation, discovery, lateral movement, command and control, and exfiltration (MITRE 2021).

Persistence should be the first post-exploitation action utilized once the cyberspace operator is able to gain initial access to a system. Persistence allows the operator to maintain long-term presence within a dynamic cyberspace environment. Once persistence is achieved, the number of malicious activities is bounded by the objective of the attacker and the defensive posture of the system.

Once persistence has been achieved, the attacker can conduct discovery actions under the disguise of a trusted user. The goal of discovery is to understand the targeted environment to support CO (MITRE 2021). Discovery can encompass accounts on the user's machines, internet bookmarks, directory services, and system information. Although user privileges can provide an abundance of information, the cyber operator may still need to perform privilege escalation for subsequent cyber actions. To conduct lateral movement, the attacker may need to conduct initial access actions, discovery, and privilege escalation post-exploitation actions within the network to achieve the desired cyber objective(s). This may include establishing an encrypted channel for command and control from the endpoint, followed by exfiltration of information (MITRE 2021). An alternate objective could be to conduct a cyberspace attack that could harm systems or data. Post-exploitation offers the attacker many different avenues to achieve desired objectives.

2.2 Post-Exploitation Frameworks

There are numerous open-source exploitation frameworks currently used to perform initial and post-exploitation actions. Although there are many niche frameworks that address specific use cases, frameworks such as Penetration Testing Execution Standard (PTES) and ATT&CK framework are agnostic of specific use cases, which can be tailored for post-exploitation actions.

PTES was created in 2009 to "provide both businesses and security providers with a common language and scope for performing penetration testing" (PTES 2014). PTES comprises two major sections: the standard and the technical guidelines. The standard has seven tailorable sections that explain how to plan, test, and report the results of a penetration test. The technical guide provides a baseline process of how to carry out different aspects of the penetration test. The post-exploitation portion consists of five tailorable sections. Although the PTES provides the base understanding of penetration testing, it requires an experienced operator to implement processes to connect the technical to the standard aspects of PTES.

The main concept of ATT&CK is to prioritize cyberspace defense based on "documented threat behavior" by understanding an adversary's tactics and techniques (Strom 2018). A tactic answers the "why" or the reason an operator performs an action. A technique is the "how" an attacker will achieve the tactic. The menu of tactics and techniques can be aggregated to create an attack scenario to match an Advanced Persistent Threat (APT), such that it can be simulated by red teamers to test and help network defenders identify the specific techniques utilized to understand where to search for compromise.

Unlike the PTES, the ATT&CK framework does not have a post-exploitation section. This is because an attacker can use both initial access and post-exploitation actions to achieve an objective. The lack of boundaries in the ATT&CK framework provides the flexibility required to track adversarial behavior from initial access to post-exploitation.

2.3 Post-Exploitation Tools

There are a variety of post-exploitation tools that are utilized by penetration testers and red teamers. Unfortunately, the DOD lacks resources to meet the required need of exploitation services. To make matters worse, the process to certify and accredit red teams "does not evaluate a red team's ability to portray validated

threat actors," so red teams "gravitate to low hanging fruit" such as known initial access exploits, which network defenders can readily identify (Schab 2021).

We researched three open-source tools that automate aspects of post-exploitation actions: Automated Network Exploitation Through Penetration Testing (ANEX), Scalable Automated Vulnerability Scanning & Exploitation Tool (SAVE-T), and Automated Deep Learning Post-Exploitation (Booz 2020; Dazet 2016; Maeda & Mamoru 2020).

These automated post-exploitation tools have a variety of limitations. First, each tool only tests a certain subset of the foundational ATT&CK cyber actions, which are discovery, persistence, privilege escalation, and lateral movement. The post-exploitation actions covered in ANEX, SAVE-T, and Automated Deep Learning tools are lateral movement and privilege escalation. SAVE-T delves into command and control and exfiltration actions, but agents are placed on target computers prior to the start of red teaming or penetration testing, circumventing the need to identify the risk associated with initial access vulnerabilities. To give network defenders a true assessment of external and internal cyberspace defenses, many threat scenarios require assessing initial access vulnerabilities along with post-exploitation actions.

Second, each implementation suffers from user misuse error. ANEX could crash the targeted system based on the type of exploit utilized. SAVE-T requires a remote access trojan (RAT) to be installed on each workstation such that if the SAVE-T server is compromised, all workstations are compromised. The deep learning tool requires the user to choose probability thresholds to train an agent, resulting in mixed assessments. Lastly, each implementation requires in-depth knowledge of the system architecture, moderate user interaction, and risk analysis of performing an automated assessment.

The U.S. Navy uses a tool called Vulnerability Remediation Asset Manager (VRAM) that gives enterprise visibility into a network's baseline configuration protecting it from known initial access vulnerabilities (Barnett 2020). VRAM splits vulnerabilities into various categories, a program-of-record owned (POR-owned) vulnerability is the most severe (SPAWAR 2017). A POR-owned vulnerability requires the owning program office to investigate and provide remediation procedures to the on-site network defenders. On-site network defenders are either left to operate with POR-owned vulnerabilities or be forced to take the associated system or component offline. To make matters worse, there could be many POR-owned vulnerabilities that need to be analyzed for their potential impact on operations. If the organization decides to operate with POR-owned vulnerabilities, dwell time is increased by the adversary executing persistence post-exploitation actions.

FireEye defines dwell time as "the number of days an attacker is present in a victim environment before they are detected." In 2021, that dwell time was 21 days (FireEye 2022). From the time a vulnerability is identified, it could take on-site network defenders 21 days to detect an adversary. The time between notification, investigation, and implementing remediation procedures allows adversaries to conduct post-exploitation actions.

If the network defenders are experienced penetration testers, they can employ post-exploitation actions to understand the impact of the exploited POR-owned initial access vulnerability. Typically, network defenders are not trained to employ post-exploitation actions and they must compete for scarce post-exploitation services. These services require a request for resources from U.S. Fleet Cyber Command, which is adjudicated with U.S. Cyber Command. The further down the chain a command is from the combatant command, the less likely the request will be fulfilled. Once a red team or penetration tester performs post-exploitation, network defenders are given a technical report outlining successful post-exploitation actions. Oftentimes these actions are performed using proprietary tools which cannot be re-enacted or reemployed to understand network effects.

An automated post-exploitation tool fills the gap between normal operations and awaiting manual penetration and red teaming services. Network defenders do not have a reliable automated tool or taxonomy that requires both minimal CO expertise and interaction that provides an insightful assessment.

3. Impact-Based Post-Exploitation Taxonomy

We implemented the most important post-exploitation actions for OCO: discovery, persistence, privilege escalation, and lateral movement. Without these foundational actions, other cyberspace post-exploitation actions like cyberspace attack, command and control, and exfiltration would not be possible. Cyberspace attack and command and control require, at a minimum, the actions of discovery to ensure that the proper system is being targeted, and persistence to maintain access to the target.

In addition to discovery and persistence, exfiltration will also require privilege escalation to use privileged user services like file transfer protocol for data movement. If the data to exfiltrate must traverse multiple networks, lateral movement will be required to establish the exfiltration path.

To understand the impact of these foundational actions, a network defender could employ an ATT&CK inspired impact-based post-exploitation taxonomy based on red team, penetration testing, and CARTT reports to assess the qualitative impact of post-exploitation actions depicted by Figure 1.

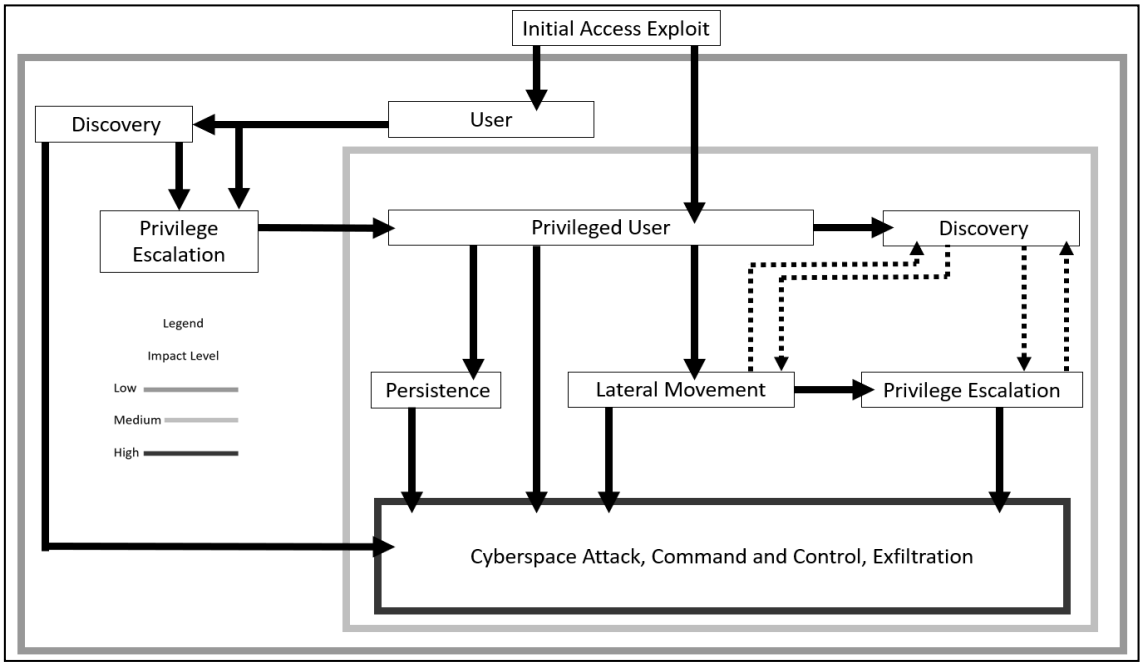


Figure 1: Impact-based post-exploitation taxonomy

The initial access exploit is the first step into either a user or privileged user system. A user system could support various business objectives which could threaten organizational security. A privileged user system provides greater access and authorization in an enterprise. Based on business impact analysis, the impact of exploiting a user or privileged user system can be defined. The impact taxonomy relies on successfully completed post-exploitation actions adapted from the ATT&CK framework. A successful completion of a post-exploitation action results in a low, medium, or high impact level. Impact level determination guidelines that could be used are the Federal Information Processing Standards Publication 199 or the Chairman Joint Chiefs of Staff Instruction 6510.01B (CJCS 2012). The impact levels can be adapted to fit any organization's security concerns of availability, integrity, and confidentiality.

If a user workstation becomes the victim of an initial access exploit, it could be considered a low impact on business operations availability if that workstation is lost. Although an adversary could disrupt business operations by taking many user workstations offline, it may or may not amount to a high-impact cyberspace attack based on business impact analysis. A user could easily switch to another workstation, or the security operations center could enact business continuity procedures to shift operations to another site, if the business impact reaches the established threshold for a cyberspace attack. If the adversary executes discovery actions to find that the workstation belongs to a high value target, they can conduct privilege escalation to either laterally move to a user system, persist on the exploited user system, or continue discovery actions at the privileged user level. If the adversary is a privileged user, they could potentially shut down the workstation denying access to business services, or laterally move into and exploit other user systems and perform shutdowns. The impact level of the adversary is raised to medium due to the scale and potential effect they could have on availability. If the adversary gained access to a privileged user system, a medium or high impact would immediately be assessed due to the ease of gaining access which could result in a cyberspace attack that results in a loss of availability for a privileged user system.

Based on the report of successful post-exploitation actions employed, the user can understand, assess impact, and mitigate risk using the impact-based post-exploitation taxonomy.

4. Design

CARTT utilizes a client-server architecture to enable users to complete cyber actions as shown in Figure 2. In the previous version of CARTT, the Operator user role was limited to vulnerability scanning and initial access exploitation, but it has now been expanded to include post-exploitation actions. CARTT leverages the Greenbone Vulnerability Management system (GVM) – formerly called OpenVAS – and the Metasploit Framework (MSF) to enable access to its expanded functionality. These tools were chosen primarily due to their open-source nature, which provides free access to the tools, and ensures real-time improvements from the open-source community. The CARTT server provides all the components required to conduct OCO.

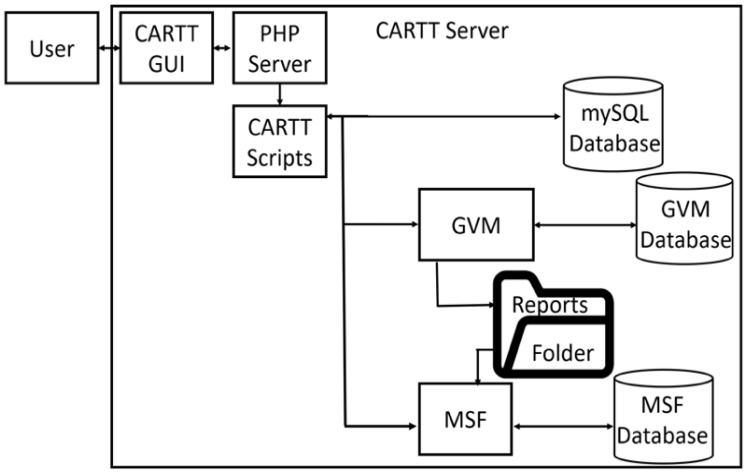


Figure 2: CARTT server architecture

The user interacts with the CARTT server via its graphical user interface (GUI) to perform cyber actions (Gomandakoye 2021). Based on the user interaction, the PHP server parses user input to create CARTT script interaction. The CARTT scripts enable communication between the following three systems: the mySQL database, GVM, and MSF. The mySQL database is used for the messaging system between users. GVM, a public vulnerability scanner maintained by Greenbone Networks, identifies vulnerabilities by conducting vulnerability scans on a system based on feeds comprised of Common Vulnerabilities and Exposures (CVEs), security communities, Greenbone labs, and input from technology partners (Greenbone 2021). The feeds are stored in the GVM database. Once a vulnerability scan is complete, its results are stored in the reports folder. From the reports folder, a scan report is imported by the CARTT Operator into MSF for exploitation actions, as shown in Figure 3, the CARTT Server process architecture.

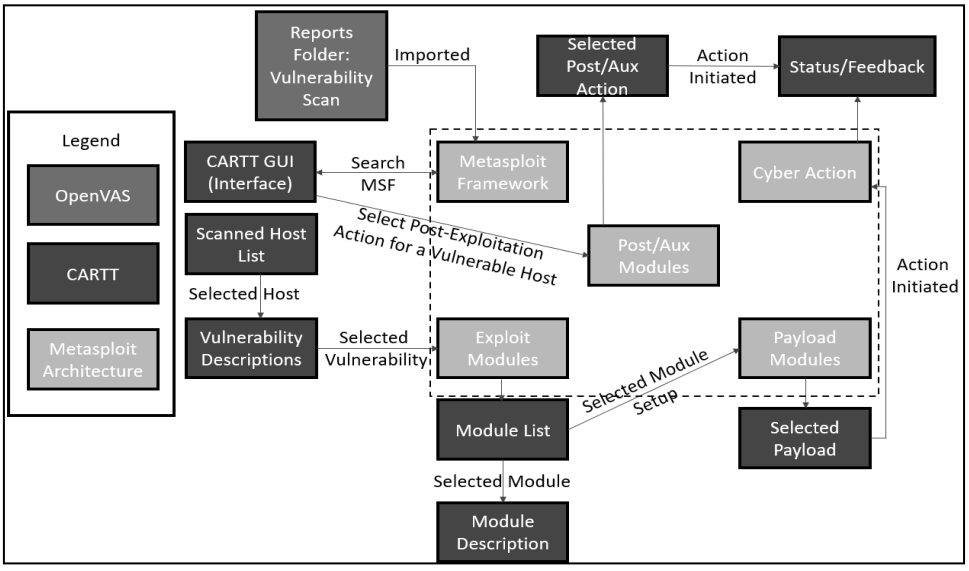


Figure 3: CARTT server processes

Each vulnerability scan is placed into a reports folder and then imported into MSF. The CARTT Operator is presented with the scanned host list in the CARTT GUI to select a host and view the related vulnerability descriptions (shown in purple in Figure 3). The Operator then selects a vulnerability, which populates a list of applicable exploit modules. The Operator can view a module description to aid with selection. Once a module has been selected, the Operator will select a payload, then verify the exploit and payload to initiate the initial access exploit cyber action. Once the initial access exploit is complete, the Operator will receive its status and feedback from the CARTT GUI. The Operator can now select and configure post-exploitation actions, as shown in Figure 3. After a post-exploitation action has been completed, the Operator can view its status and any system feedback on its success or failure.

When a user logs into CARTT as an Operator, they are directed to a webpage to enter the target IP and subnet to conduct a new vulnerability scan. Once the scan is completed, the Operator can import the completed scan and then begin target exploitation from a list of available target hosts. The Operator will then submit a selected Target Host IP, which prompts CARTT to provide a list of vulnerabilities for that target. After selecting an available vulnerability by its CVE, the Operator will be provided a list of MSF exploit modules from which to launch an initial access exploit. CARTT presents the results of this exploitation to the Operator, who then has the option to use the same or a different exploit to conduct post-exploitation actions.

In post-exploitation, the CARTT Operator can choose from among several ATT&CK post-exploitation actions or test all actions. Upon selecting a post-exploitation action, the Operator is directed through various workflows that require minimal input and interaction to complete the cyber action. Once a post-exploitation workflow has completed, the Operator will be presented its results, as in Figure 4.



Figure 4: Post-exploitation results display

In Figure 4, the results show that the initial access exploit was successful since a Meterpreter payload session was opened. In this example, the Operator chose to use the privilege escalation post-exploitation action. The ATT&CK technique for conducting privilege escalation leveraged *process injection* using the MSF ***post/windows/escalate/getsystem*** module (highlighted in Figure 4), which exploits a race condition on the target system.

After execution of the MSF module ***post/windows/escalate/getsystem*** the results are provided to the CARTT Operator, showing the successful execution of privilege escalation. Based on this, one could refer to the impact-based post-exploitation taxonomy discussed in Section 3 to determine whether the initial access vulnerability, if exploited, would result in a medium-level impact to the organization. The organization could then implement appropriate security controls to mitigate the risk impact.

5. Implementation

This section discusses the virtualized environment used for implementing the new post-exploitation capabilities in CARTT. It also describes the resource scripting and workflows used for implementing specific post-exploitation actions.

5.1 Environment

The implementation environment used for CARTT is shown in Figure 5. The machines used for implementation were virtualized through VMware Workstation Pro. The environment was divided into two networks: Network-A depicting an operational environment, and Network-B a developmental environment. The operational environment is used to perform business tasks, while the development environment is used to update the products that users utilize to complete business tasks. A CARTT operator could move laterally between the two environments to complete a CO objective.

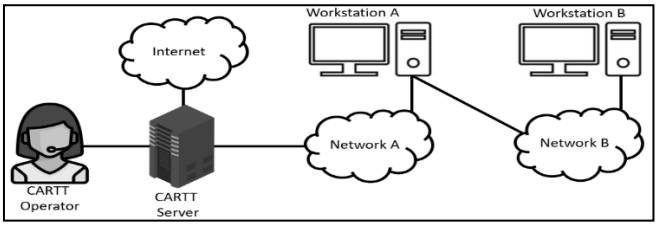


Figure 5: Implementation environment

The CARTT server is connected to Network-A and to the internet, whose connectivity is required to utilize GVM capabilities. Workstation-A in Network-A is dual-homed, since it has an additional connection to Network-B. Workstation-B is connected only to Network-B. The additional workstation was used to implement and test automation of the lateral movement post-exploitation action. For testing privilege escalation, which was the post-exploitation action example discussed in Section 4, CARTT gained access to Workstation-A using an initial access exploit, and then executed the privilege escalation post-exploitation action. All post-exploitation actions of discovery, persistence, privilege escalation, and lateral movement were developed, implemented, and successfully tested using this implementation environment.

5.2 Resource Scripting

CARTT leverages Metasploit resource scripting to allow MSF interoperability with the PHP web-based CARTT GUI. To implement post-exploitation actions, MSF post modules were appropriately added and logically chained in various Metasploit resource scripts and PHP webpages. The resource script enables automation by aggregating multiple Metasploit commands. The code for the initial access exploit resource script is shown in Figure 6.

Code

```
$fd_rc = fopen("user_data/exploit_$user.rc", 'w+');

$script =
"$exploit\n
set RHOST $host\n
set ExitOnSession false\n
exploit -z\n
exit -y\n";

fwrite($fd_rc, $script);
fclose($fd_rc);

$cmd = "msfconsole -q -r
user_data/exploit_$user.rc";
exec($cmd);
```

Figure 6: MSF initial access resource script

A file named *exploit_$user.rc* is the resource script file created using the C library function *fopen*, which is stored in the variable named *$fd_rc*. The *$fd_rc* will store the *$script* variable. The *$script* variable holds the set of Metasploit commands to complete the initial access exploit action. The *$exploit* is the initial access exploit that is set for use. The *set RHOST $host* is the IP address of the target. The *ExitOnSession* determines if session state will be maintained across multiple instances of Metasploit (Rapid7 2021). If the *ExitOnSession* option is set to true, and a Meterpreter session is obtained, the Meterpreter session will be terminated upon exit. If *ExitOnSession* is set to false, the Meterpreter session state is maintained upon exit.

To implement post-exploitation, *ExitOnSession* is set to false to maintain the open Meterpreter session to allow CARTT to execute post-exploitation actions. The *exploit -z* command runs the exploit and if a Meterpreter session is obtained, the session is placed in the background and returns to the MSF console environment. The *exit -y*

MSF command will exit the MSF environment. The *$script*, which holds the set of Metasploit commands to complete initial access exploitation is written to $fd_rc. The exploit_$user.rc resource script file is then executed through the Metasploit command line interface using the command *exec($cmd)* where *$cmd = msfconsole -q -r user_data/exploit_$user.rc*.

To implement post-exploitation actions, MSF post modules were appropriately added to the Metasploit resource script. It is important to note that the initial access exploit must succeed prior to performing any post-exploitation actions.

5.3 Workflows

Each post-exploitation menu option directs the CARTT Operator to a different webpage based on a PHP script that runs the menu. The webpage will prompt the Operator for a set of inputs to progress through a series of webpages that will eventually lead to the results of the post-exploitation action. The inputs are captured by the PHP POST method that stores the information in a PHP SESSION variable used across post-exploitation webpages (Chan 2020).

The CARTT post-exploitation actions of discovery, persistence, and privilege escalation follow a common workflow to arrive at providing results to the Operator, as shown in Figure 7 below. Each box in the workflow represents a different PHP webpage, and based on the CARTT Operator input, will either provide feedback or validate input prior to transition to the next webpage. The Operator must provide the required input to receive the results of the post-exploitation action.

The CARTT Operator will then click one of the post-exploitation actions of discovery, persistence, or privilege escalation, which will direct them to the "Configure Initial Access Exploit" page to provide the Vulnerability, Module, and Local Host IP Address. Upon submission, the action is completed, and results are displayed to the CARTT Operator (refer to Figure 4).

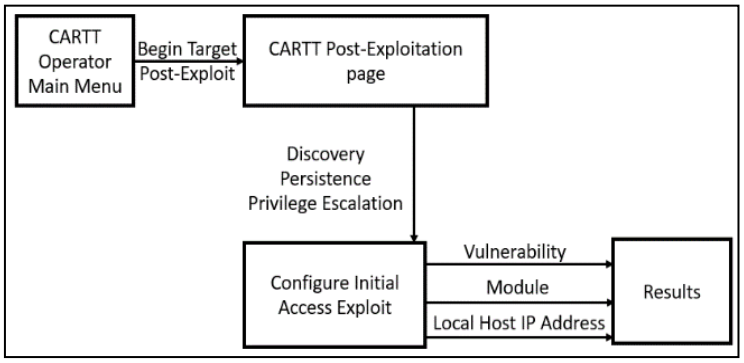


Figure 7: Post-exploitation workflow

The post-exploitation actions to perform lateral movement, or to test all post-exploitation actions, require two additional webpages to gather the required input from the Operator, as shown in Figure 8 below. If *Lateral Movement* is selected, the Initial Access Exploit Results will contain the results from executing a custom resource script CARTT MSF discovery module. If *Test All Actions* is selected, the Initial Access Exploit Results will also include the results of the *Persistence* and *Privilege Escalation* actions.

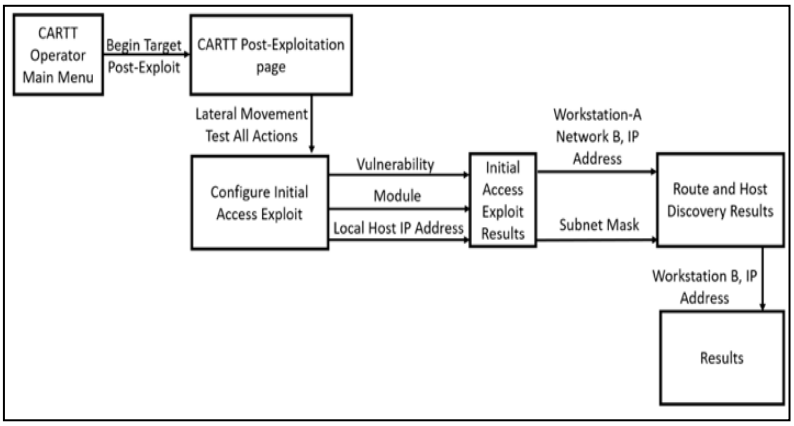


Figure 8: Lateral Movement and Test All Actions workflows

The CARTT Operator will use the custom resource script CARTT MSF discovery module results to provide the Workstation-A, Network-B IP address as well as the subnet mask. The Operator will then be presented the Route and Host Discovery Results and will be prompted to input the Workstation-B IP address. Upon submission the Operator will be presented with the results.

6. Conclusions and Future Work

This research expanded CARTT to support post-exploitation cyber actions by adding appropriate MSF modules to Metasploit resource scripts to perform discovery, persistence, privilege escalation, and lateral movement on a target host. CARTT uses MSF resource scripting coupled with PHP operability to automate the new post-exploitation actions. This capability provides the CARTT Operator with valuable information beyond initial access exploitation. CARTT also provides organizations with the capability to conduct self-assessment and impact-based analysis when higher-level cyberspace analysis resources are not available. Once organizations understand both external and internal impacts, they can harden their systems to increase their cybersecurity posture.

In future work, CARTT can be made more effective by expanding it to model obfuscation, stealth, and non-attribution to better model threat behaviors. To demonstrate the new post-exploitation capabilities, CARTT was tested against workstations that were not fully hardened, however, most real-world target organizations will have host and network intrusion and prevention devices implemented behind firewalls or within a DMZ. Another aspect to improve upon is automating initial access exploitation. If a CARTT Operator is unfamiliar with specific target vulnerabilities, this may require them hours of research to pair a vulnerability with an exploit module to achieve initial access. Finally, CARTT reporting and user feedback could be improved. The current system reporting requires the user to manually parse the results of a vulnerability analysis to find the proper inputs to complete a post-exploitation workflow. Future work could parse and present information in a cleaner manner to provide improved feedback and action suggestions to the user.

References

Barnett, J. (2020) “New cybersecurity tools sought for Navy’s VRAM program”, [online], FEDSCOOP, https://www.fedscoop.com/navy-cybersecurity-vram-cots/

Booz, J. (2020) “Towards scalable automated vulnerability scanning & exploitation”, [online], M.S. Thesis, Dept. of Info. Sci., Carnegie Mellon University, USA, https://kilthub.cmu.edu/articles/thesis/Towards_Scalable_Automated_Vulnerability_Scanning_Exploitation/12728360/1?file=24095777

Benito, R. (2022) “An automated post-exploitation model for offensive cyberspace operations”, [online], M. S. thesis, Naval Postgraduate School, Monterey, CA, USA, https://calhoun.nps.edu/bitstream/handle/10945/70631/22Jun_Benito_Ryan.pdf.

Berrios, J. (2020) “A client/server model for automated red teaming”, [online], M. S. thesis, Naval Postgraduate School, Monterey, CA, USA, https://calhoun.nps.edu/bitstream/handle/10945/66584/ 20Dec_Berrios_Joseph.pdf.

Chairman Joint Chiefs of Staff (2012) “CJCSM 6510.01b: Cyber incident handling program”, [online], https://www.jcs.mil/Portals/36/Documents/Library/Manuals/m651001.pdf?ver=2016-02-05-175710-897.

Chan, J. (2020), *Learn PHP in One Day and Learn It Well*. Hedge End, Hampshire, UK: LCF Publishing.

Dazet, E. (2016) “ANEX: automated network exploitation through penetration testing”, [online], M.S. Thesis, Dept of Comp. Sci., California Polytechnic State University, San Luis Obispo, USA, https://digitalcommons.calpoly.edu/theses/1592/.

Department of Defense (2021) “Department of Defense (DOD) zero trust reference architecture,” [online], https://dodcio.defense.gov/Portals/0/Documents/Library/(U)ZT_RA_v1.1(U)_Mar21.pdf.

FireEye (2022) “M-trends 2022”, [online], https://www.mandiant.com/resources/m-trends-2022.

Gomandakoye, O.M. (2021) “An expanded graphical user interface for web-based cyber automated red teaming tool (CARTT)”, [online], M. S. thesis, Naval Postgraduate School, Monterey, CA, USA, https://calhoun.nps.edu/bitstream/handle/10945/68326/21Sep_Goumandakoye_Ousmane.pdf.

Greenbone, “Security feed: The daily vulnerability update.” Accessed May 5, 2021 [Online]. Available: https://www.greenbone.net/en/security-feed/

Maeda, R. and Mamoru, M. (2020) “Automating post-exploitation with deep reinforcement learning”, *Comp. & Sec.*, vol.100, pp. 1-13, January, [online], https://doi.org/10.1016/j.cose.2020.102108.

MITRE, “ATT&CK matrix for enterprise,” Accessed July 31, 2021 [Online]. Available: https://attack.mitre.org/

NIST (2021) “Federal information processing standards publication standards for security categorization of federal information systems”, [online], https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf.

PTES (2014), “Penetration testing execution standard,” Accessed July 14, 2021 [Online]. Available: http://www.pentest-standard.org/ index.php/Main_Page

Rapid7, “Meterpreter HTTP/HTTPS communication,” Accessed April 28, 2021 [Online]. Available: https://www.rapid7.com/blog/post/2011/06/29/meterpreter-httphttps-communication/

Rose, S., Oliver, B., Stu, M., and Connelly, S. (2020) “NIST special publication 800-207 zero trust architecture”, [online], https://csrc.nist.gov/publications/detail/sp/800-207/final.

Schab, J. (2021) “Tackling DOD cyber red team deficiencies through systems engineering”, [online], SANS Institute, Bethesda, MD, USA, https://sansorg.egnyte.com/dl/Bx3K1e25qH/?.

Space and Naval Warfare Systems Center (2017) “Vulnerability remediation asset manager 2.0: VRAM quick start guide”.

Strom, B. (2018) “The philosophy of ATT&CK”, [online], https://medium.com/mitre-attack/the-philosophy-of-att-ck-9e8f81aba119.

Review of End-to-End Encryption for Social Media

Vijay Bhuse
School of Computing and Information Systems, Grand Valley State University, Allendale, MI, USA

bhusevij@gvsu.edu

Abstract: People have valid concerns about their privacy and the use of their personal information by corporations. People do not necessarily trust social media companies to protect their right to privacy. Social media companies are under pressure to provide greater levels of security and privacy to their users. The current gold standard of security protocols for messaging system is the Signal Protocol. The Signal protocol is an open-source end-to-end encryption model. It uses AES-256, HMAC-SHA256 and Curve25519 as its cryptographic primitives. This protocol is currently considered cryptographically sound and provides excellent information security. However, many social media companies are still using less secure protocols often underpinned by less secure primitives. This paper discusses in detail the various cryptographic primitives used in social media apps like WhatsApp, Twitter, Facebook, Snapchat and Instagram.

Keywords: Encryption, Cryptanalysis, Privacy, Signal Protocol, end-to-end encryption, public key cryptography.

1. Introduction

Social Media has revolutionized global communication and continues to influence it. Social media apps have become integrated into our everyday lives. Due to concerns over Cambridge Analytica's use of Facebook data and a subsequent movement to encourage users to abandon Facebook, there is a renewed focus on how social media companies collect personal information and make it available to marketers, as per Rainie (2018). An incredible amount of personal information is posted on these platforms or is sent through them via their messaging. Social media has proved to be a vital communication tool for protestors and political dissidents in various parts of the world. Primarily due to these factors, social media companies have been under increasing pressure from privacy advocates to implement stronger and stricter data security measures.

In response, several social media apps began to utilize end-to-end encryption via the Signal Protocol. While end to end encryption is a secure device to device communication model, the cryptographic primitives underpinning it are vital to its security. Karbasi (2021) states that the Signal Protocol originally utilized AES-256, HMAC-SHA256 and Curve25519, all three of which are considered cryptographically secure.

As per WhatsApp (2020), WhatsApp implemented full end-to-end encryption via the Signal Protocol in 2016. Its parent company, Facebook, implemented optional end-to-end encryption in Facebook Messenger, named Secret Conversations. Facebook has also announced that they are working to merge Facebook Messenger, Instagram Direct Message, and WhatsApp into a single message service with default end-to-end encryption. Wyrich (2020) states that Twitter has yet to deploy any end-to-end encryption for direct messages, however, the company has publicly considered the idea. As per Twitter (2021) there is scarce publicly available information on the exact protocols and primitives. Twitter currently uses secure direct messages. Snapchat is similar, the company claims that all pictures sent by users are end-to-end encrypted. Snapchat has not released any details on the protocols or primitives used. Any chats sent via Snapchat are not end-to-end encrypted and again, no technical details on how they are secured have been released as per Amnesty (2016).

This paper is organized as follows. We discuss end-to-end encryption in section 2. We discuss the cryptographic primitives used for end-to-end encryption in section 3, and end with discussion and conclusions in section 4.

2. End-to-end Encryption

First, a brief overview of what end-to-end encryption is and why it is considered cryptographically secure. As per Greenberg (2014), End-to-end encryption is a way for two processes on two different devices to send messages to each other in such a way that it is impossible for a third party, even one that is relaying the messages between the two, to view them. The basic principle behind it is the use of public key cryptography and a secure key transfer between devices. This will only allow those two processes to decrypt any messages sent between them.

The Signal Protocol is an open-source implementation of end-to-end encryption. Signal is a secure messaging application that uses end-to-end encryption to protect the privacy of its users. It has been widely recognized for the strength of its security features and has been endorsed by privacy advocates and security experts, including

Edward Snowden, who is a well-known whistleblower and privacy advocate. Signal is considered one of the most secure messaging apps available, due to its use of end-to-end encryption and other security measures such as forward secrecy. It is also open source, which means that its code can be independently audited for security vulnerabilities. As per Amnesty (2016) International, many activists, journalists, and political dissidents have recommended Signal as a safe and secure way to communicate with others, as it provides a high level of protection against government surveillance and other forms of digital spying.

The Signal Protocol uses a combination of public key cryptography and symmetric key cryptography to provide end-to-end encryption for messages and calls. The Signal Protocol uses the Elliptic Curve Diffie-Hellman (ECDH) key agreement protocol, specifically the Curve25519 curve, to establish a shared secret key between two devices. The private key is used to create a public key, which is shared with the other device. Once the shared secret key is established, the Signal Protocol uses it to generate three session keys: one for encrypting messages, one for authenticating messages, and one for encrypting calls. These session keys are used to encrypt and authenticate messages and calls using AES-256 and HMAC-SHA256, respectively. AES-256 is widely considered to be a very secure encryption algorithm and it is used by many organizations and governments worldwide. HMAC-SHA256 is a widely used method for message authentication, it creates a digest of the message and a secret key, that is sent along with the message, the recipient uses the same key to authenticate the message. Overall, the Signal Protocol is designed to provide a high level of security and privacy for its users, and it is considered one of the most secure messaging protocols available today, as per Karbasi (2021).

As per WhatsApp (2020), the Signal Protocol uses a ratchet mechanism to ensure forward secrecy, which means that if the encryption keys of one device are compromised, the attacker will not be able to decrypt past messages. During the initial key exchange, the devices use Elliptic Curve Diffie-Hellman (ECDH) key agreement to establish a shared secret key for each pair of public keys. These shared secrets are then combined to create a master secret key. Each device then derives a Root Key and a Chain Key from the master secret key. When a message is sent, the sender's device uses the current Chain Key to derive a Message Key, which is used to encrypt the message. The sender's Chain Key is then "ratcheted" forward, and a new Chain Key is derived. This ensures that the Message Key used to encrypt the current message cannot be used to derive any previous Chain Keys. Additionally, the Signal Protocol periodically creates new Root Keys and Chain Keys through a new Elliptic Curve Diffie-Hellman agreement. This ensures that even if the Root Key of a device is compromised, the attacker will not be able to derive the Chain Keys used to encrypt current and future messages. All these mechanisms ensure that the Signal Protocol provides forward secrecy, which is an important security feature that prevents past conversations from being decrypted if a device's keys are compromised later.

All the generated key pairs, apart from the Identity Key, are temporary, and are recreated frequently to limit the amount of data compromised should the keys leak. When these keys are regenerated, the key exchange process happens again, and a new master secret is created, as per Karbasi (2021).

From a typical user's perspective, one of the downsides of the Signal Protocol is that it does not support cross-device message transfer. Each device generates its own key pair and is responsible for encrypting and decrypting messages. This means that messages sent or received on one device cannot be decrypted on another device, even if the devices are owned by the same user. This means that if a user wants to use Signal on multiple devices, they will have to manually transfer their conversations and contacts to each device and would not be able to seamlessly switch between devices. This approach has been chosen to maximize the security and privacy of the users, keeping the private keys on the device and not allowing any third party to access them, thus ensuring that if one device is lost or compromised, the private keys and messages cannot be accessed. However, Signal is working on a feature called "Sealed Sender" that will allow cross-device messaging and syncing, and still maintain the same level of security and privacy.

3. Cryptographic Primitives

The cryptographic primitives that underpin end-to-end encryption are vital to its security. The Signal Protocol uses the following cryptographically secure primitives.

3.1 Curve25519

Curve25519 is the fastest elliptic curve that can be used in the Elliptic Curve Diffie-Hellman (ECDH) key exchange. The ECDH key agreement protocol allows two parties to establish a shared secret key over an insecure communication channel. It creates a shared secret from the two users using one user's private key and the other's public key. This shared secret is then used by both parties to encrypt and authenticate their messages between each other. Elliptic curve cryptography shared secret keys are significantly smaller than standard shared secret keys, while retaining comparable cryptographic strength. This difference grows rapidly as the key sizes increase. An elliptic curve key with a length of 512 bits is equal to a standard asymmetric key of 15,360 bits. The smaller key size also increases the speed of the key exchange, as per Niasar (2020). The use of Curve25519 in the Signal protocol contributes to the overall security and performance of the protocol, by providing a fast and secure key agreement mechanism that can be used to establish shared secrets between devices.

3.2 AES-256

AES (Advanced Encryption Standard) is a widely used symmetric encryption standard that refers to a set of block ciphers. The ciphers are differentiated by their key size, with AES-256 (256 bit key) being the most used. AES encryption is performed in a series of rounds, during which the data is substituted, permuted, and shifted. The number of rounds depends on the key size, with AES-256 requiring 14 rounds.

AES is widely considered to be a secure encryption standard, and it is approved for use on state secrets by the United States government. AES is also considered to be immune from brute force attacks with current technology.

The Signal Protocol uses AES-256 to encrypt messages between clients. This provides a high level of security for the messages, as AES-256 is a very secure encryption algorithm. WhatsApp and Facebook Messenger's Secret Conversations use AES-256-CBC to encrypt messages. CBC (Cipher Block Chaining) is a mode of operation for block ciphers, it makes the encryption more robust against certain types of attacks, however it doesn't include any message authentication, which is why both platforms use HMAC-SHA256 for that as per Kishore (2016). Overall, the use of AES-256 in combination with the CBC mode of operation and HMAC-SHA256, provides a high level of security for messages sent through these platforms.

3.3 HMAC-SHA256

HMAC-SHA256 is a hash algorithm used for authentication and to verify message integrity. HMAC stands for hash-based message authentication code. SHA-256 refers to the hash algorithm used in the creation of the authentication code. The security of the authentication code is dependent on the underlying hash function. SHA-256 is a specific hash algorithm in the SHA-2 standard that is considered cryptographically secure at this time, as per Kishore (2016).

When a user goes to send a message, the shared secret, created during the Elliptic Curve Diffie-Hellman Exchange, is used to derive two keys that will be hashed with the message. HMAC-SHA256 hashes the combination of the first key and the message. It then hashes the combination of the resulting hash and the second key. The result is a unique combination that can only be recreated by the exact same combination of the message, and the keys derived from the shared secret, as per WhatsApp (2020).

This primitive is the only one of the three that is currently in danger of no longer being considered cryptographically secure. The security of its predecessor, SHA-1 is questionable. SHA-2 algorithms were a refinement of SHA-1 algorithms, but the two share the same algorithmic architecture. The fear is that improvements of the techniques used to compromise SHA-1 could compromise SHA-2. Currently, this has yet to occur, and is a purely theoretical fear, as per Kishore (2016).

3.4 Other Primitives Used

Not all social media platforms utilize the Signal Protocol, but all use some sort of encryption to safeguard user data from outside attacks as well as provide authentication. Twitter developers have released many APIs designed to integrate or provide access to Twitter into various other media platforms. A common API is the button on an article or video that allows a user to easily share the content to their Twitter.

These APIs obviously must be able to securely handle user data. However, Twitter has released extraordinarily little technical data on exactly how they do this. Twitter API documentation merely says that the APIs use the TLS cipher suite to secure communications. It would be logical to assume that Twitter uses similar methods to secure the direct messages of users, but no further details are provided, as per Twitter (2021).

The TLS cipher suite comprises of some block ciphers and some stream ciphers, along with supported hash algorithms to be used for message authentication and data integrity. Not all the supported algorithms are considered cryptographically secure. It is unlikely that Twitter would use a depreciated encryption algorithm, but the lack of transparency on the matter raises concerns.

Similarly, Snapchat has made no technical details about message encryption public. The company has stated that it does use end-to-end encryption for pictures sent by users but provided no details on primitives. It does not utilize end-to-end encryption for messages. In a response to a query from Amnesty International on their security protocols, Snapchat only stated that they utilize transport encryption and encrypt user data on both their servers and the user's device. In 2016, Snapchat announced that they were working on a secure messaging system, but they informed that they cannot share any specific details, as per Amnesty (2016) International.

4. Discussion and conclusions

From a cryptographic primitive standpoint, Facebook-owned message platforms utilize some of the strongest primitives among the platforms discussed. The Signal Protocol is a highly secure encryption protocol, and WhatsApp uses it as the default encryption method for end-to-end encryption.

Twitter and Snapchat also use encryption, but they have not published the details of their encryption protocols, encryption schemes, or cryptographic primitives, making it difficult to evaluate the level of security they provide.

End-to-end encryption is the most secure way to encrypt messages between two devices. It ensures that the service carrying the message is unable to view it, providing a high level of privacy. But it can also be used by criminals to hide their activities, making it a "double-edged sword" with valid arguments on both sides.

Overall, the use of strong cryptographic primitives, such as the Signal Protocol, can help ensure the security and privacy of messages sent through messaging platforms. However, it is important for companies to be transparent about the encryption methods and algorithms they use, so users can make informed decisions about the apps they use.

References

Amnesty (2016), "For Your Eyes Only? Ranking 11 Technology Companies on Encryption and Human Rights." Amnesty International, 2016.

Greenberg, A., (2014). "Hacker Lexicon: What Is End-to-End Encryption?". *WIRED*, Archived from the original on 23 December 2015.

Karbasi, H., Amir, and Shahpasand, S. (2021) "SINGLETON: A Lightweight and Secure End-to-End Encryption Protocol for the Sensor Networks in the Internet of Things Based on Cryptographic Ratchets." *The Journal of Supercomputing: An International Journal of High-Performance Computer Design, Analysis, and Use* 77(4): 3516.

Kishore, N., and Kapoor, B., (2016) "Attacks on and Advances in Secure Hash Algorithms." *IAENG International Journal of Computer Science* 43(3): 25–34.

Niasar, M., El Khatib, R., Azarderakhsh, R., and Mozaffari-Kermani, M. (2020). "Fast, Small, and Area-Time Efficient Architectures for Key-Exchange on Curve25519." In *2020 IEEE 27th Symposium on Computer Arithmetic (ARITH)*, , 72–79.

Orlic, V., Peric, M., Banjac, Z., and Milicevic., S. (2012) "Some Aspects of Practical Implementation of AES 256 Crypto Algorithm." *2012 20th Telecommunications Forum (TELFOR), Telecommunications Forum (TELFOR), 2012 20th*: 584–87.

Twitter (2021). "Security." *Twitter - Documentation - Security*. https://developer.twitter.com/en/docs/security.

Rainie, L. (2018), "Americans' complicated feelings about social media in an era of privacy concerns", https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/, Archived from the original on 27 March 2018.

Whatsapp, (2021), "WhatsApp Encryption Overview: Technical White Paper". https://cryptome.org/2016/04/whatsapp-crypto.pdf, Archived from the original on April 2016.

Wyrich, A. (2020). "Twitter's Hack Reveals Glaring Security Concerns around DMs." *The Daily Dot*. https://www.dailydot.com/debug/twitter-end-to-end-encryption-dms/.

An Analysis of Crypto Scams during the Covid-19 Pandemic: 2020-2022

Johnny Botha[1], Danielle Petronella Botha[2] and Louise Leenen[3]
[1,2]Council for Scientific and Industrial Research, Pretoria, South Africa
[3]University of Western Cape and CAIR, Cape town, South Africa

jbotha1@csir.co.za
dbotha3@csir.co.za
lleenen@uwc.ac.za

Abstract: Blockchain and cryptocurrency adoption has increased significantly since the start of the Covid-19 pandemic. This adoption rate has overtaken the Internet adoption rate in the 90s and early 2000s, but as a result, the instances of crypto scams have also increased. The types of crypto scams reported are typically giveaway scams, rug pulls, phishing scams, impersonation scams, Ponzi schemes as well as pump and dumps. The US Federal Trade Commission (FTC) reported that in May 2021 the number of crypto scams were twelve times higher than in 2020, and the total loss increased by almost 1000%. The FTC also reported that Americans have lost more than $80 million due to cryptocurrency investment scams from October 2019 to October 2020, with victims between the ages of 20 and 39 represented 44% of the reported cases. Social Media has become the go-to place for scammers where attackers hack pre-existing profiles and ask targets' contacts for payments in cryptocurrency. In 2020, both Joe Biden and Bill Gates' Twitter accounts were hacked where the hacker posted tweets promising that for all payments sent to a specified address, double the amount will be returned, and this case of fraud was responsible for $100,000 in losses. A similar scheme using Elon Musk's Twitter account resulted in losses of nearly $2 million. This paper analyses the most significant blockchain and cryptocurrency scams since the start of the Covid-19 pandemic, with the aim of raising awareness and contributing to protection against attacks. Even though the blockchain is a revolutionary technology with numerous benefits, it also poses an international crisis that cannot be ignored.

Keywords: Blockchain, Covid-19, Crypto-crime, Cryptocurrency, Crypto-scams.

1. Introduction

Blockchain adoption has increased significantly in recent years, and so has crypto-crime. Some may argue that Bitcoin gives power back to the people. However, the popularity of cryptocurrencies in recent years has attracted the attention of many scammers and fraudsters. Undoubtedly, the rise of cryptocurrency has added to the immense increase in crime rates. Illicit transactions in cryptocurrency have reached a staggering $14-billion in 2021, an 80% increase from 2020 - which constitutes a new record. Chainalysis[1], a blockchain data platform, stated that the rise of decentralised finance (DeFi), is the leading factor behind the spread of crypto scams (Chainalysis, 2021). Out of all hacks and scams in 2021, 21% took advantage of a loophole in DeFi (Xia, et al., 2020).

Scammers have been around long before crypto, but some of the characteristics of crypto are very appealing to them. Crypto has no middleman as in the case with banks. Instead, direct transactions occur between two individuals. The Covid-19 pandemic brought difficult times with job losses and salary cuts. People became desperate to invest in alternative methods and crypto seemed like the perfect solution with the consequence that scammers took advantage of this opportunity. Crimes in the cryptocurrency space victimise innocent people - it places a big barrier in its further adoption and increases government restrictions. Investigating and exploring cryptocurrency transactions remain intractably hard due to its pseudonymous nature and with every cryptocurrency having its own protocol and blockchain (Social Links, 2022).

This study analyses the most significant crypto-scams since the start of the Covid-19 pandemic. It aims to raise awareness and contributes towards protection against these attacks. Blockchain is a revolutionary technology with immense benefits. However, the technology attracts crime and poses an international crisis.

2. Types of Crypto Crimes and How to Avoid Them

2.1 Giveaway

One of the most common scams is the giveaway scam. This is where the attacker lures the victim in by announcing to give away certain cryptocurrencies or assets. For example, one instance is where a contribution

[1] https://www.chainalysis.com

of one Ethereum coin (sent to a specified address) results in double the amount returned. Normally a target will receive a link to a landing page website, with fake transactions indicating that participants are indeed getting paid. In several cases the website has a timer that is activated with the intent of placing the target under the impression that there is a limited time for the offer to be taken up (Bureau, 2022).

Several social media sites are being used for these scams such as YouTube, Twitter and Instagram. In the case of YouTube, livestreams of interviews with celebrities are available. The stream links are surrounded by text details on how to participate in these schemes. Furthermore, it will also appear as if thousands of people are participating in the livestreams. However, these are generally bots and not real people. The YouTube account will often also appear to have been verified. In these cases, the accounts were hacked, all contents were deleted, and the attackers run their own livestreams. Other ways of launching this type of attack is when an attacker runs advertisements that appear around legitimate videos. Targets will then click on these advertisements and be transferred to a fake site.

2.2 Rug Pull

A new scam called a "rug pull" refers to the expression "someone pulls the rug from underneath you". They are also referred to as exit scams. This type of scam appears to be an investment company where people can invest funds with very attractive returns. The company will consequently disappear, and victims are robbed of their funds. Back in 2017, this was mostly driven by the Initial Coin Offering (ICO) schemes. Today they are more prevalent in the Non-Fungible Tokens (NFTs) and Decentralised-Finance (DeFi) spaces.

In the case of NFTs, the scammer creates a collection of images, issues them to be minted and promises an exciting roadmap with great returns. Once the mint process is completed and all people have procured their NFTs, the project creators delete the website and all social media linked to the project, thus pulling the rug, and leaving their victims upended. Another method is called the slow rug pull. In this case, the creators slowly extricate themselves from the project over several weeks or even months, causing targets to lose interest in the project. The slow rug pull is more common as this leaves no real traces of intent to defraud pointing to the attacker.

The DeFi space differs from NFTs and ICOs. Instead of sending money to a project or protocol, one needs to supply liquidity. The liquidity is used in a decentralised exchange and investors can get very lucrative returns. The creators will hype the project to increase its demand and to increase the liquidity pool size. After people have staked funds for some time on the DeFi platform, the creators will withdraw all the coins from the liquidity pool, removing all the value injected into the currency by investors, driving the price very low or even to down zero in some cases. This is called a DeFi rug pull and is the most common exit scam. Three types of rug pulls exist within DeFi:

- *Liquidity Stealing*: Token creators extract all coins from the liquidity pool driving the price to zero.
- *Limiting sell orders*: The developers are the only ones who can sell tokens. As soon as they received investments they sell the tokens for other currency pairs, leaving a worthless token to investors.
- *Dumping*: Developers immediately sell off enormous numbers of tokens, reducing the price and leaving investors with worthless tokens.

According to Moody (2022), the number of rug-pulls has been increasing annually since 2020. In 2021 the biggest financial losses of over $2.28 billion due to rug pulls have been recorded. Figure 1 presents the total amount of cryptocurrency (in USD) stolen in rug-pulls during 2020 to August 2022. Note that the graphs in Figures 1, 3 and 4 were compiled by the authors based on data from Moody (2022).

The data compiled in Figure 1 only captures rug pulls from 2020 until August 2022. During the remainder of 2022, it is important to note that further rug pulls may take place which could increase the total value lost. Many scams that took place in previous years, were only identified weeks or months after it took place. This may also be the case regarding the current data. Data capturing events taking place after the writing of this paper can be accessed via Moody (2022).

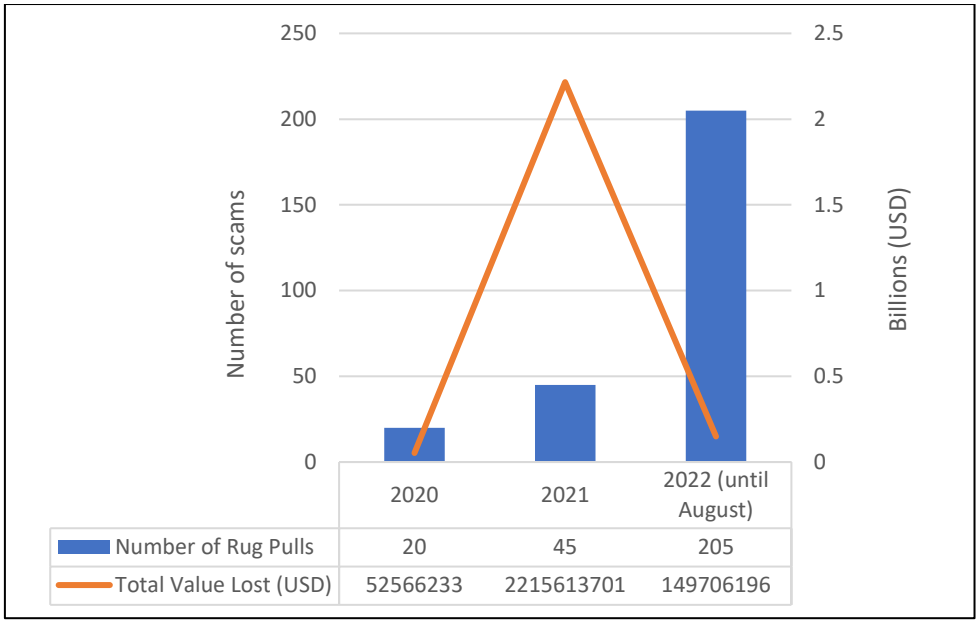


Figure 1: Total cryptocurrency value stolen in rug pulls vs the number of rug pulls

2.3 Phishing Scams

Phishing scams are pervasive within any system where human gullibility or inattention can be used against the user. As blockchain and Web 3.0 are evolving, so too are the techniques employed by phishers (Andryukhin, 2019). Scammers are fooling people to log into fake cryptocurrency exchange websites, getting access to their exchange account details and stealing their funds. Attackers have also started to directly target people's cryptocurrency wallets. The most damaging scam is when an attacker obtains the user's wallet private keys. This can happen by making users believe they need to reset their passwords and provide their secret recovery seed phrases. Another method is to ask the user to enter the seed phrase to access a certain website to allow a connect to their wallet. Once the attacker has this seed phrase, he has full access to your wallet (Bureau, 2022).

2.4 Impersonation

Impersonation is when the scammer tries to take advantage of some famous person's trustworthy reputation on social platforms such as Twitter, Telegram, Discord, Instagram, TikTok, etc. These scammers will send a direct message (DM) on the platform, pretending to be a celebrity and to offer advice to the victim on some crypto investment. They will provide a number asking the victim to WhatsApp them, where they will act as if they are providing more personal assistance. They will explain how the target can double their money by sharing a link on which the victim will click and send money to the platform or crypto address. This scam continues after the victim has sent the money; the attacker will send a fake proof of profits generated. The attacker will then request a withdrawal fee or pay them their cut of the profits first (Bureau, 2022).

2.5 Ponzi Scheme

Ponzi schemes are fraudulent investments or scams, promising high rates of returns, which do occur in the initial phases. It is similar to a pyramid scheme and normally mostly benefits early investors. Returns are only sustainable by bringing in more investors (Chen, 2021). The scheme makes people believe that a cloud mining package, lending scheme or very successful trading bot provide daily profits. These schemes also rely on current members referring their friends and family for additional rewards, which is called multi-level marketing (MLM) schemes (Bureau, 2022).

2.6 Pump-and-Dump

A pump-and-dump is when insiders "pump" or increase a token's price until a point where it creates attention and market interest. The moment others jump in, the initial investors will "dump" or sell their coins, causing a massive decrease in price, leaving late investors at a loss. A pump will be set to happen at a particular date and time. Figure 2 shows an example of such a pump-and-dump in a timespan of only three hours. Note that by the time the coin to pump is announced, the value has already been increased. These schemes are well coordinated in places such as Telegram groups. As soon as the public buys in, the organisers will do the dump causing losses for late investors.



Figure 2: Pump-and-Dump (Kamps, 2018)

3. The Worst Crypto-Crimes during the Pandemic

Using data consolidated by Moody (2022), it is clear that the number of Crypto scams has risen annually. Scams such as Ponzi Schemes, impersonations, Rug Pulls, Exit Scams, Phishing and Pump-and-Dumps were considered, whilst money laundering has been excluded. While the data available for 2022 is incomplete at the time of writing, the number of scams has already surpassed previous years, as seen in Figure 3.

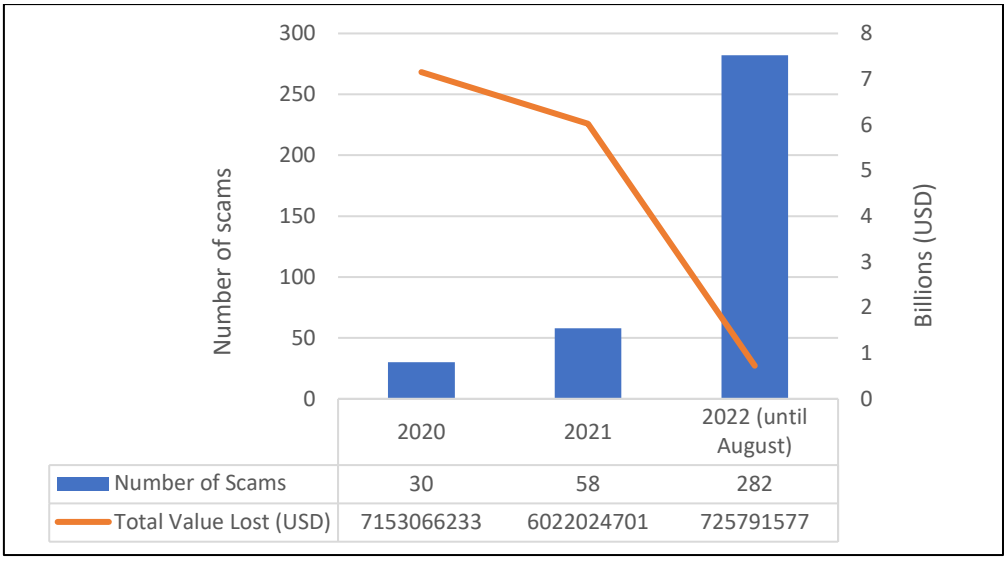


Figure 3: Number of scams per year with corresponding financial losses

Interestingly, the total value lost in 2022 is less than in previous years. However, more scams may still come to light during the remainder of the year; it is important to note that the total value lost to scams may still increase. Large-scale crashes within the crypto space, such as the Luna network (Forbes, 2022) and FTX exchange (Berwick, 2022) are both recent occurrences that, if confirmed to be scams after investigation, could push the total value lost in 2022 up by $60Bn and $2Bn respectively. This would make 2022 the year accruing the biggest financial losses.

A breakdown of the two major scam types based on monetary losses, Ponzi Schemes and Rug-Pulls, are presented in Figure 4, alongside all other losses in scams such as phishing, pump-and-dumps, impersonation,

etc. Note that some crypto scams of 2022 are suspected to be Ponzi schemes, but this has yet to be confirmed. It is possible that the amount lost to Ponzi schemes could be greater as more evidence is found.

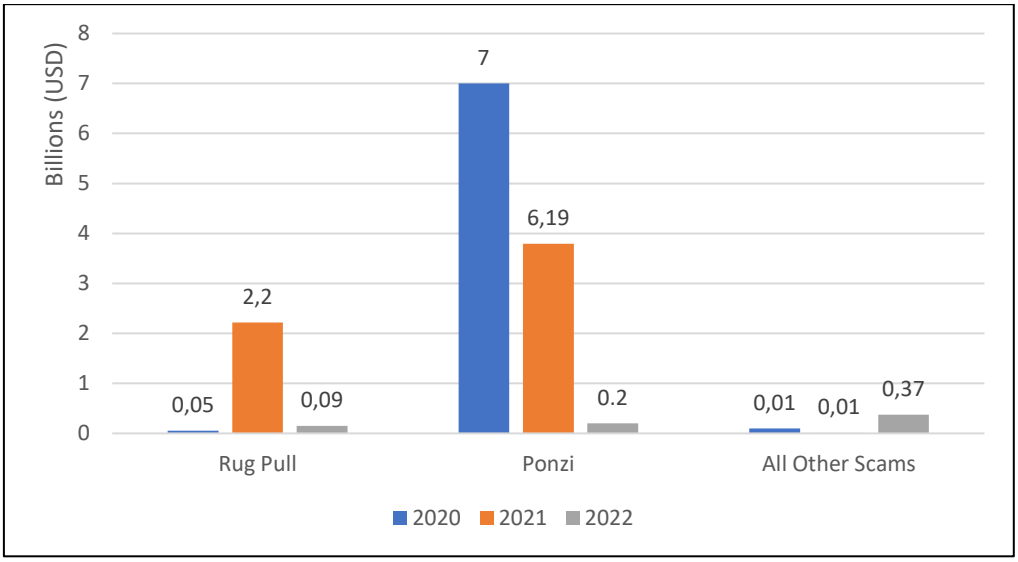


Figure 4: Financial Losses by Scam Type per Year

Table 1 lists the fifteen largest of these scams from 2020-2022, based on financial losses. The scams listed in the table are discussed below in more detail, grouped according to the year in which they occurred. In addition, the scams that affected the most users are investigated, as well as the global landscape of occurrences. Lastly, arrests made in relation to the scams are discussed.

In 2020, at the advent of the pandemic, approximately 40% of the largest scams recorded in Table 1 took place, consisted completely of Ponzi schemes. The largest of these schemes, PlusToken, was a multinational pyramid scheme whose operators have since been indicted. This scheme promised astronomical returns on investment and incentivised current investors to recruit new members, a trademark Ponzi scheme tactic (Leng, 2020). The racket owned roughly 1% of the Bitcoin supply at the height of the operation (Harper, 2020). Closely following PlusToken in losses in 2020 is MTI; The scheme accepted buy-ins in Bitcoin and is the largest fraudulent scheme that has been charged by the CFTC (Commodity Futures Trading Commision, 2022). Arbistar 2.0 SL (Nelson, 2020), BitClub (Crypto-Mining-Pool) (United States Department of Justice, 2022), and Forsage (Securities and Exchange Comission, 2022a) all offered similar returns on investments and were indicted as Ponzi or Pyramid Schemes.

At the start of 2021, the largest scams seemed to diversify, including several high-profile rug-pulls alongside more Ponzi Schemes. The Africrypt rug pull, considered one of the largest-ever crypto heists alongside PlusToken, is currently involved in an ongoing investigation and liquidation approved by South-African courts (Zimwara, 2021). Thodex, a cryptocurrency exchange, used aggressive campaigns such as luxury cars to lure investors. In April 2021, Thodex was shut down, having stolen approximately $2 billion (Baltrusaitis, 2022). Other Ponzi Schemes in this timeframe include Finiko (Tassev, 2021) and EmpiresX (Securities and Exchange Commission, 2022b), both affecting hundreds to thousands of users by offering too-good-to-be-true returns on investment. In addition, four noteworthy rug-pulls took place; Anubis Decentralised Autonomous Organisation (DAO) (Hakki, 2021), DeFi100 (Kolhatkar, 2021), Meerkat Finance (Sopov, 2021) and Snowdog DAO (Radmilac, 2021). All these rug-pulls were DeFi-based offerings that "rugged" investors after draining liquidity pools before delivering a real product. Some (such as Meerkat Finance and Africrypt) initially placed the blame on hacking events that never took place.

Table 1: The most significant crypto scams during the pandemic 2020-2022

Name	Type	Year	Losses (financial)	Country of origin	Affected users	Arrests made
PlusToken	Ponzi Scheme	2020	$4 billion	China & South-Korea	±3 million	Yes
Africrypt	Rug Pull	2021	$3.6 billion	South-Africa	±2 million	No
Thodex	Rug Pull	2021	$2 billion	Turkey	±390,000	Yes
MTI	Ponzi Scheme	2020	$1.7 billion	South-Africa	±280,000	Yes
Arbistar	Ponzi Scheme	2020	$1 billion	Spain	Up to 32,000	Yes

Name	Type	Year	Losses (financial)	Country of origin	Affected users	Arrests made
BitClub	Ponzi Scheme	2020	$722 million	Netherlands	Unspecified-thousands	Yes
Forsage	Ponzi Scheme	2020	$300 million	Russia & USA	Unspecified-millions	Yes
Morris Coin	Rug Pull	2022	$140 million	India	±1.1 million	Yes
Ormeus Coin	Ponzi Scheme	2022	$124 million	USA	±12,000	Yes
EmpiresX	Ponzi Scheme	2021	$100 million	USA	Unspecified-thousands	Yes
Finiko	Ponzi Scheme	2021	$95 million	Russia	Unspecified-hundreds	Yes
AnubisDAO	Rug Pull	2021	$60 million	Hong-Kong	Unknown	No
DeFi100	Rug Pull	2021	$32 million	Unspecified	Unknown	No
Meerkat Finance	Rug Pull	2021	$31 million	Unspecified	Unknown	No
SnowdogDAO	Rug Pull	2021	$30 million	Unspecified	Unknown	No

During 2022 (August at the time of writing), fewer noteworthy scams have been uncovered. So far, two large rug-pulls are Morris Coin and Ormeus Coin. The former was a fake Initial Coin Offering (ICO) (Jacob, 2022) and the latter a coin purportedly to be used as a mining currency (Pimentel, 2022). Both lured investors via social media.

Though some scams are difficult to trace due to anonymity extended via blockchain technologies (ByBit Learn, 2021), the location of origin could be determined for most of the top scams. The distribution of the scams is quite even spread globally, with South-Africa, Russia and the USA being the countries producing multiple of the most significant scams. Notably, South-Africa produced two of the top four crypto scams, as seen in Table 1.

PlusToken (Leng, 2020), Africrypt (Zimwara, 2021) and Morris coin (Jacob, 2022) affected the most users. Generally, Ponzi Schemes seem to affect more users. This could be due to the business model – investors are incentivised to recruit new members in the hope of increasing their profits. Though the Morris Coin scam did not cause financial losses as large as the Ponzi schemes, it amassed many participants (Jacob, 2022). Further, it can be difficult to quantify the number of people affected by rug-pulls. Sellers may obfuscate sales progress to hinder attempts to ascertain the legitimacy of the project (Investopedia, 2022).

Roughly 66% of the discussed cases resulted in arrests. Individuals were charged in various manners, including accounts of fraud, market manipulation, and more. The scams that have not yet resulted in arrests are mostly rug pulls. Creators often remain anonymous, making it much harder to track bad actors (ByBit Learn, 2021).

4. Protection Against Crypto Scams

4.1 Giveaway and Impersonation

The giveaway scam is normally linked to the impersonation scam where the scammer pretends to be a legitimate platform such as Coinbase[2] promoting a 5000 BTC giveaway when you send them, for example, 1 BTC. A legitimate investment platform or exchange will never ask for crypto in exchange for receiving more in return. Other type of impersonations are celebrities and YouTube live streams. These impersonator accounts will often have more followers than the real account to confuse the users, but these followers are all bots. To avoid these scams, awareness should be raised that no famous influencer on YouTube, Twitter, TikTok, etc. will send a DM to some unknown individual. Celebrities have thousands of followers and do not have the time to interact with followers one-on-one. Platforms such as Twitter and Telegram gives a blue tick for verified accounts. In addition, the public should not click on links sent by someone that offers crypto investment advice.

The only way to avoid these types of scams is to understand that no-one on the Internet is going to give something away for free, and no one will double an investment amount. If it sounds too good to be true, it normally is. One should think twice before sending crypto funds because all transactions are irreversible and participants will not be able get their money back (Hauer, 2020; Bureau, 2022).

4.2 Rug Pull

Rug pulls are more common in new cryptocurrency projects and investors should make sure they are choosing established projects. For example, Bitcoin has been used worldwide and its inner workings have been reviewed

[2] https://www.coinbase.com

thoroughly. Newer projects do not have such a track record and leave room for hiding certain aspects from investors. One indication, although it is not a guarantee, is to establish whether a new cryptocurrency is listed on well-known exchanges such as Binance[3] or Coinbase[2]. These exchanges do thorough reviews on assets before listing them. However, the trade-off is that the highest rewards may come from projects before they are listed on these exchanges. Scammers normally prey on the fear of missing out (FOMO) on massive gains, luring investors towards projects before being listed on well-known exchanges (Rosen, 2022).

To avoid a crypto project rug pull one would need to perform broad due diligence. Research needs to be done before investing and extremely high annual percentage yield (APY) promises should heighten caution (Bureau, 2022). Before investing, an investor should understand how a product works and not just blindly invest based on hype. Investors should also determine if the company has been registered and does indeed exist. In addition, establishing if a project has been audited (refer to LCX audits by CERTIK[4]) and doing research on the project team are advised. If no actual person can be identified behind a project and only pseudo-names are being used, it is a red flag (Rosen, 2022).

If there are any limits on sale orders, it is an indication of a scam project. If no liquidity is locked and the price explodes with a small number of token holders, it should raise some questions. It is recommended to perform transaction and on-chain data analysis. It is most likely a rug pull if there is little trading activity and it appears only in a few decentralised exchanges (DEXs). Potential investors should stay calm and avoid FOMO. Excessive advertising content indicates a rug pull as most new coins normally starts slow and small. Other common signs to look out for are if the project appeared overnight, developers are anonymous, there is a low liquidity, the total value locked (TVL) is low, the liquidity is unlocked, disproportionate token distribution, low effort website and a lack of social media presence (Vardai, 2022).

In the case of NFTs, one should be a lot more discerning when buying or minting. Only 5% of NFT projects will be very profitable and 95% would probably go to zero. This type of investment is a very high risk, and one should do a lot of research before participating (Bureau, 2022).

4.3 Phishing Scam

An investor should never enter his seed phrase anywhere on a website to access the website or to send any funds. Leaving a seed phrase on your personal computer or in the cloud is a very risky. It has been reported that hackers have stolen $655K from one MetaMask[5] user by picking the seed from an iCloud backup. To avoid a seed phrase from being included in a backup, one should exclude the MetaMask (hot wallet) app from iCloud backups via the iOS settings. In addition, enabling two-factor authentication for MetaMask is advised. It is also recommended to keep cryptocurrencies in a cold wallet if they are not actively being traded. Keeping investments out of social media would make an individual less of a target as hackers are keeping an eye out for high-value victims (Toulas, 2022).

Another attack method is when the attacker gives a victim their seed phrase on a fraudulent site. When the victim installs the wallet and enters the seed thinking it is their seed, when in fact it is under the scammer's control. Once the victim sends funds to the wallet, the scammer runs off with it. Sometimes scammers will buy advertisement space on Google for their links to be visible before the actual real website is available, enticing users to click on these links and follow their installation steps (Akhtar, 2021). To avoid these scams, never click on any advertisements for crypto wallets. The attacker can try to make the victim connect their wallet to a fraudulent decentralised application (DApp), signing a smart contract and allowing the attacker to spend on their behalf. This scam is popular under the NFT DApps. An example of this occurred in 2021 where a Bored Ape Yacht Club[6] collector lost 16 high value NFTs, valued around $2.2 million, when he approved a phishing contract (Chawla, 2021). To avoid this type of scam, a user must make sure they are on the official DApp website and that they know which DApps they have approved. Etherscan[7] provides the functionality to show all the smart contracts a user has given approval to and when it was given. It also allows a user to revoke certain smart contracts that the user does not recognize.

[3] https://www.binance.com/en
[4] https://www.certik.com/projects/lcx
[5] https://metamask.io
[6] https://opensea.io/collection/boredapeyachtclub
[7] https://etherscan.io

4.4 Ponzi Scheme

To avoid a Ponzi scheme, one needs to be very sceptical and compare the returns to the market average. If the scheme is consistently above that average, it should raise suspicion. In addition, one should be able to find out exactly how an investment platform is generating returns (Bureau, 2022). The US SEC (Securities and Exchange Commission, 2022c) released a report detailing "red-flags" to help identify possible crypto Ponzi schemes - presented in Table 2.

Table 2: SEC Ponzi Scheme Red-Flags

Red flag	Description
Low-risk, high-returns	If an investment guarantees exceedingly attractive, high returns on investment, one should be very sceptical.
Overly consistent returns	If an investment does not fluctuate with market conditions, it could be an indicator of a Ponzi scheme.
Investments that are unregistered with an authority	If an investment is not regulated by some authority, it can be exceedingly difficult to determine the legitimacy thereof.
Unlicenced Sellers	Many Ponzi schemes consist of unlicensed sellers. If a seller is registered, it could indicate involvement with Ponzi schemes.
Overly complex fee structures	If the fee structures are too complex or secretive, it can be obscuring true intentions. Structures that are too complex to understand or are incomplete, should be avoided.
Minimum investor qualifications not necessary	Legitimate investment opportunities afforded to individuals generally require the investor to be accredited. If an investment does not require your net worth or salary, one should be sceptical.
Paperwork issues	If an investment opportunity is slow to produce paperwork/makes excuses as to why paperwork is unavailable in writing, one should be sceptical.

4.5 Pump and Dump

Pump-and-dumps are illegal and should always be avoided. Firstly, they occur in low market cap coins because it is easier to move the prices of these coins. Secondly, the coins will mostly be on shady exchanges. Thirdly, if there is no news about a certain coin's price increase, it is more likely the works of scammers. One can also look at previous trading volumes; if a token has had low trading volumes for the past few months and suddenly a spike occurs, it could be a clear sign of early accumulation. The pump operators need to buy the coins upfront before they dump them, therefore the accumulation waves need to be assessed (Bureau, 2022).

Pump-and-dump schemes are normally more based on hype and speculation than on a business model. They tend to create a heightened sense of urgency to invest. If a company is not yet profitable, common sense should be used as in to why one would invest. Potential investors should look at a company's financial statements and earnings report (Beers, 2022).

5. Conclusion

Blockchain adoption continues to increase and so does crypto-crimes and scams, especially since the Covid-19 pandemic. With job losses and salary cuts, people became desperate for a new source of income. The crypto space provided this opportunity, but it also created the perfect playground for scammers. This study highlights the most significant crypto-scams from 2020 to 2022, based on monetary losses and the number of users affected. The scams have been grouped according to type, the year they occurred, financial losses, country of origin, number of affected users and whether any arrests have been made. In addition, the study aims to raise awareness on protection against these scams. Recommendations are provided on identifying a specific scam and how to avoid them. Since there is no official legislation in place for crypto-scams, it remains a very significant international threat that cannot be ignored.

References

Akhtar, T. (2021, 11 4). *MetaMask, Phantom Wallet Users Targeted in Crypto Phishing Scam: Report*. Retrieved from Coindesk.com: https://www.coindesk.com/business/2021/11/04/crypto-phishing-scam-targets-metamask-phantom-wallet-users-report/

Andryukhin, A. (2019). Phishing Attacks and Preventions in Blockchain Based Projects. *International Conference on Engineering Technologies and Computer Science (EnT)* (pp. 15-19). Moscow: IEEE.

Baltrusaitis, J. (2022, September). *Thodex crypto exchange's CEO arrested for allegedly defrauding investors of $2.7 billion.* Retrieved from Finbold: Cryptocurrency News: https://finbold.com/thodex-crypto-exchanges-ceo-arrested-for-allegedly-defrauding-investors-of-2-7-billion/

Beers, B. (2022, 09 22). *Pump-and-dump scheme: What it is and how to avoid one*. Retrieved from Bankrate.com: https://www.bankrate.com/investing/pump-and-dump-scheme/

Berwick, A. (2022, November 14). *At least $1 billion of client funds missing at failed crypto firm FTX*. Retrieved from Reuters: https://www.reuters.com/markets/currencies/exclusive-least-1-billion-client-funds-missing-failed-crypto-firm-ftx-sources-2022-11-12/

Bureau, C. (2022, 5 18). *WORST Crypto Scams in 2022!! DONT Fall For These!!* Retrieved from YouTube: https://www.youtube.com/watch?v=GKTa5ciCJl4

ByBit Learn. (2021, August 27). *Why Crypto Rug Pulls Happen in DeFi — and How to Avoid Them*. Retrieved from ByBit Learn: https://learn.bybit.com/investing/why-crypto-rug-pulls-happen-in-defi/

Chainalysis. (2021). *The 2021 Crypto Crime Report.* Chainalysis. Retrieved from https://go.chainalysis.com/rs/503-FAP-074/images/Chainalysis-Crypto-Crime-2021.pdf

Chawla, V. (2021, 12 30). *Bored Ape NFT Collector Loses $2.2M in Phishing Scam*. Retrieved from Cryptobriefing.com: https://cryptobriefing.com/bored-ape-nft-collector-loses-2-2m-in-phishing-scam/

Chen, J. (2021, 10 21). *Ponzi Scheme*. Retrieved from Investopedia: https://www.investopedia.com/terms/p/ponzischeme.asp

Commodity Futures Trading Commision. (2022, June 30). *CFTC Charges South African Pool Operator and CEO with $1.7 Billion Fraud Involving Bitcoin*. Retrieved from CFTC.gov: https://www.cftc.gov/PressRoom/PressReleases/8549-22

Forbes. (2022, September 20). *What Really Happened To LUNA Crypto?* Retrieved from Forbes Digital Assets: https://www.forbes.com/sites/qai/2022/09/20/what-really-happened-to-luna-crypto/?sh=302d8e074ff1

Hakki, T. (2021, October 31). *AnubisDAO Investors Lose $60 Million in Alleged Rug Pull*. Retrieved from Decrypt: https://decrypt.co/84924/anubisdao-investors-lose-60-million-in-alleged-rug-pull

Harper, C. (2020, August 19). *How The PlusToken Scam Absconded With Over 1 Percent Of The Bitcoin Supply*. Retrieved from Bitcoin Magazine: https://bitcoinmagazine.com/business/how-the-plustoken-scam-absconded-with-over-1-percent-of-the-bitcoin-supply

Hauer, T. (2020, 4 6). *Crypto giveaway scams and how to spot them*. Retrieved from blog.Coinbase.com: https://blog.coinbase.com/crypto-giveaway-scams-and-how-to-spot-them-59e24d220616

Investopedia. (2022, March 28). *How to Identify Cryptocurrency and ICO Scams*. Retrieved from Investopedia: https://www.investopedia.com/tech/how-identify-cryptocurrency-and-ico-scams/

Jacob, S. (2022, January 8). *Morris Coin scam: At least 1.1 million investors may have been cheated*. Retrieved from Business Standard: https://www.business-standard.com/article/companies/morris-coin-scam-at-least-1-1-million-investors-may-have-been-cheated-122010800045_1.html

Kamps, J. K. (2018). To the moon: defining and detecting cryptocurrency pump-and-dumps. *Crime Science 7* (p. Article 18). Crime Science Journal. doi:https://doi.org/10.1186/s40163-018-0093-5

Kolhatkar, S. (2021, May 24). *DeFi100 Coin Scam: DeFi100 Creators Rug Pull Investors, Disappear With $32 Million*. Retrieved from Republic World: https://www.republicworld.com/technology-news/other-tech-news/defi100-coin-scam-defi100-creators-rug-pull-investors-disappear-with-32-dollars-million.html

Leng, S. (2020, December 1). *Chinese cryptocurrency scam ringleaders jailed in US$2.25 billion Ponzi scheme involving PlusToken platform*. Retrieved from China Macro Economy: https://www.scmp.com/economy/china-economy/article/3112115/chinese-cryptocurrency-scam-ringleaders-jailed-us225-billion

Moody, R. (2022, September 20). *Worldwide crypto & NFT rug pulls and scams tracker*. Retrieved from Comparitech: https://www.comparitech.com/crypto/cryptocurrency-scams/

Nelson, D. (2020, October 22). *Spanish Police Arrest Head of Billion-Dollar Crypto Arbitrage Platform on Fraud Allegations*. Retrieved from CoinDesk: https://www.coindesk.com/markets/2020/10/22/spanish-police-arrest-head-of-billion-dollar-crypto-arbitrage-platform-on-fraud-allegations/

Pimentel, B. (2022, March 8). *Two siblings were charged in a global $124 million crypto fraud operation*. Retrieved from Protocol: https://www.protocol.com/bulletins/ormeus-crypto-fraud

Radmilac, A. (2021, November 27). *Avalanche's first memecoin SDOG ends in a $30M possible rugpull*. Retrieved from Cryptoslate: https://cryptoslate.com/avalanches-first-memecoin-sdog-ends-in-a-30m-rugpull/

Rosen, A. (2022, 6 30). *How to Avoid 'Rug Pulls,' the Latest Cryptocurrency Scam*. Retrieved from Nerdwallet.com: https://www.nerdwallet.com/article/investing/rug-pull

Securities and Exchange Comission. (2022a, August 1). *SEC Charges Eleven Individuals in $300 Million Crypto Pyramid Scheme.* Retrieved from U.S. SEC Press Release: https://www.sec.gov/news/press-release/2022-134

Securities and Exchange Commission. (2022b, June 30). *SEC Charges Empires Consulting Corp. with Fake Trading Scheme*. Retrieved from U.S. SEC: https://www.sec.gov/news/press-release/2022-119

Securities and Exchange Commission. (2022c). *Ponzi schemes Using virtual Currencies.* Office of Investor Education and Advocacy. SEC. Retrieved from https://www.sec.gov/files/ia_virtualcurrencies.pdf

Social Links. (2022, April 22). *Enhancing Cryptocurrency Investigations with OSINT*. Retrieved from https://blog.sociallinks.io/: https://blog.sociallinks.io/cryptocurrency-investigations/

Sopov, V. (2021, April 03). *Largest Binance Smart Chain Fraud: Meerkat Finance (MKAT) Rug Pulled with $32 Million Losses*. Retrieved from U.Today: https://u.today/largest-binance-smart-chain-fraud-meerkat-finance-mkat-rug-pulled-with-32-million-losses

Tassev, L. (2021, November 12). *Finiko Fugitives Suspected of Moving 750 BTC From Crypto Pyramid's Wallet*. Retrieved from Bitcoin.com: https://news.bitcoin.com/finiko-fugitives-suspected-of-moving-750-btc-from-crypto-pyramids-wallet/

Toulas, B. (2022, 4 18). *Hackers steal $655K after picking MetaMask seed from iCloud backup*. Retrieved from BleepingComputer.com: https://www.bleepingcomputer.com/news/security/hackers-steal-655k-after-picking-metamask-seed-from-icloud-backup/

United States Department of Justice. (2022, March 24). *Nevada Man Admits Money Laundering and Tax Offenses Related to BitClub Network Fraud Scheme*. Retrieved from U.S. Attorney's Office: District of New Jersey: https://www.justice.gov/usao-nj/pr/nevada-man-admits-money-laundering-and-tax-offenses-related-bitclub-network-fraud-scheme#:~:text=NEWARK%2C%20N.J.%20%E2%80%93%20A%20Nevada%20man,U.S.%20District%20Judge%20Claire%20C.

Vardai, Z. (2022, 09 22). *How to avoid getting rug-pulled in crypto and DeFi*. Retrieved from forkast.news: https://forkast.news/how-avoid-getting-rug-pull-scammed-crypto-defi/

Xia, P., Wang, H., Luo, X., Wu, L., Zhou, Y., Bai, G., . . . Liu, X. (2020). Don't Fish in Troubled Waters! Characterizing Coronavirus-themed Cryptocurrency Scams. *2020 APWG Symposium on Electronic Crime Research (eCrime)*, (pp. 1-14). doi:10.1109/eCrime51433.2020.9493255

Zimwara, T. (2021, August 3). *South Africa Bitcoin Heist: Court Grants Liquidators Authority to Track Missing Africrypt Investor Funds*. Retrieved from Bitcoin.com: https://news.bitcoin.com/south-africa-bitcoin-heist-court-grants-liquidators-authority-to-track-missing-africrypt-investor-funds/

Case Study: Conducting a Risk Assessment for an Electrical Utility

Edwin Covert, CISSP-ISSAP, CISM, CRISC, SCF, PMP
Colorado State University Global, Aurora, Colorado, USA

edwin@edwincovert.com

Abstract: Risk management is an important part of effective cybersecurity. This paper presents a hypothetical risk assessment as a case study for one of the largest electricity providers in the southern California region using the approach outlined in the National Institute of Standards and Technology Special Publication 800 series on cybersecurity.

Keywords: Risk Management, Risk Assessment, Case Study

1. Introduction

Electrical generation is the lifeblood of the US and the global economy, according to Webb (2016). One firm dominates the creation of electricity in southern California. This paper will discuss various aspects of cybersecurity risk management relevant to a local utility company's (referenced as either company or utility in henceforth) management, including how to document specific risks it might face. Next, this paper will address high-risk concerns through the application of proven mitigation techniques. As a business, the company has standard cybersecurity concerns with corporate networks, such as access control, data loss prevention and business continuity. However, what makes this firm unique is that it manages a large electrical grid. The company recognizes that it possesses critical information technology systems, sensitive customer and employee data, network infrastructure, and information vital to its business.

2. The Company's Organizational Risks

Electrical generation requires Industrial Control Systems (ICS) to create and distribute power where it needs to go. Within the utility sector, risks to ICS remain high (Khodabakhsh et al., 2020;). ICS systems are those that are "used to control industrial processes such as manufacturing, product handling, production, and distribution" (National Institutes of Standards and Technology, 2019, para. 1). According to Khodabakhsh et al. (2020), they comprise "assets such as supervisory control systems and data acquisition (SCADA), distributed control systems (DCS), and human-machine interfaces (HMI) that are commonly used for monitoring and control of their critical infrastructure" (p. 1). Vendors increasingly developed them to use internet-based communication protocols (Ovaz Akpinar & Ozcelik, 2018).

The Cybersecurity and Infrastructure Security Agency (CISA) (2021) has released detailed advisories surrounding ICS prompted by several examples of ICS attacks that made headlines. One of the first occurred in 2000 when Vitek Boden used stolen radio equipment and drove to 46 radio-controlled sewage control systems in Queensland, Australia (Abrams & Weiss, 2008). In each case, he used the stolen radio equipment to shut off industrial controls remotely that the maintenance company installed for remote access. This led to 800,000 gallons of untreated sewage released into the water supply. Citing the Australian Environmental Protection Agency, Abrams and Weiss (2008) note, "Marine life died, the creek water turned black, and the stench was unbearable for residents" (p. 1).

A second example involves Programmable Logic Controllers (PLC) and a type of malware called a worm. PLCs are another type of ICS. According to Kaspersky (2021), threat actors design worms with an eye towards self-replication: "[w]orms do not require activation—or any human intervention—to execute or spread their code" (para. 1). A worm is a self-replicating program designed to disrupt operations. Spenneberg et al. (2016) documented as a proof-of-concept a worm in a popular PLC that would consume PLC resources and significantly affect operations. Researchers presented this scenario at a renowned cybersecurity conference.

A final example of an in-the-wild risk to ICS is WIN32/Industroyer. This malware targets ICS used in electrical substations. According to Cherepanov (2017), WIN32/Industroyer targets four distinct ICS protocols, showing the threat actors had advanced knowledge of ICS design. The malware installs a separate backdoor to additional attacks, wipes data, and contains additional payloads designed to control ICS (Cherepanov, 2017).

Threats to ICS are increasing (Radanliev et al., 2018). Therefore, organizations must determine their ICS risk; for this, there is a standard formulation: risk equals the likelihood of a threat exploiting a vulnerability and causing an impact on operations (National Institute of Standards and Technology, 2012). Figure 1 shows this graphically. This paper will explore each of these as they apply to the company.

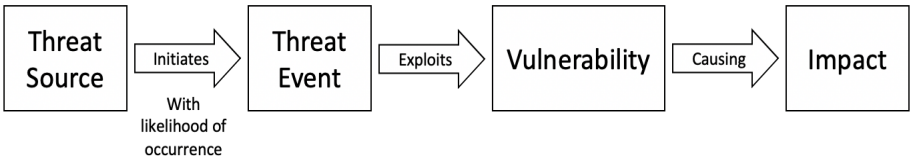


Figure 1: Determining risks

Note: Adapted from "Guide for Conducting Risk Assessments (SP 800-30, rev. 1)" by National Institute of Standards and Technology, 2012.

2.1 Threats

The National Institute of Standards and Technology (NIST) (2012) defines a threat as "any circumstance or event with the potential to harm organizational operations and assets, individuals, other organizations, or the Nation through an information system via unauthorized access, destruction, disclosure, or modification of information, and/or denial of service" (p. 8). Applied to this organization, threats come from a variety of sources. The US Intelligence Community (IC) states cybersecurity attacks from "nation states and their surrogates will remain acute" (Office of the Director of National Intelligence, 2021, p. 20). The IC notes that cybersecurity operations are increasingly a tool that countries apply to increase their national power; these operations will affect civilians and non-military targets. The Annual Threat Assessment from the IC also calls out threat actors who have attacked "software and [information technology] service supply chains" (p. 21).

While nation state threat actors are important to counter, threats come in a variety of flavors, both intentional and accidental. Additional intentional threats include criminals and activists who might target the company; the techniques might be the same but the motivations are different. Nation states seek to cause disruptions to rivals; criminals seek financial gain; activists look to increase the visibility of their cause. The company must account for accidental threats, including natural disasters and errors caused by users as well.

An important way to understand what threats the utility faces is via a threat model. A threat model creates a picture of the overall attack surface it presents to the outside world. The attack surface is what a threat actor sees at the perimeter of a system where they will try to enter (National Institutes of Standards and Technology, 2019). Understanding the attack surface is knowing what parts of the company's information systems are vulnerable and need to be tested, according to OWASP (2021).

Developing a threat model involves decomposing the information system into its individual components, identifying the risks presented (because of design or software elements used) and developing countermeasures to each of the risks identified (Shostack, 2014). Analysts use the threat modeling process to "analyze potential attacks or threats, and can also be supported by threat libraries or attack taxonomies" (Xiong & Lagerström, 2019, p. 56). Analysts then collate each of these attacks into a sense of situational awareness, i.e. the attack surface of the systems in question.

Using the STRIDE model allows organizations to consider the range of threats it faces. Each letter in the mnemonic stands for a distinct threat tactic: *S* for spoofing, *T* for tampering, *R* for repudiation, *I* for information disclosure, *D* for denial of service, and *E* for elevation of privileges. Table 2 defines each of these terms. Each intentional threat actor in this section can and will use one or more of these threat tactics to reach their aim. MITRE (2020) has created a detailed list of attack techniques threat actors have employed that the company can use to inform its threat model.

Table 2: Definitions for each letter in the STRIDE threat model method

Term	Definition
Spoofing	Pretending to be something the user or system is not
Tampering	Modifying something in an unauthorized manner
Repudiation	Falsely claiming a user or system did or did not do something
Information Disclosure	Revealing information meant to be protected
Denial of Service	Preventing a service that suppose to run from happening
Elevation of Privileges	Doing things a user or system should not be able to do

Note: Adapted from "Threat Modeling: Designing for Security" by A. Shostack, 2014. Copyright 2014 by Wiley.

2.2 Vulnerabilities

Cybersecurity professionals define a vulnerability as a "weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source" (National Institute of Standards and Technology, 2012, p. 9). There are a multitude of vulnerabilities in ICS the company must know. NIST published its *Guide to Industrial Control Systems (ICS) Security* in 2015. Appendix C of this NIST document breaks down vulnerabilities into five categories: architecture and design, configuration and maintenance, physical access, software development, and communications and network (Stouffer et al., 2015).

2.2.1 Architecture and Design

Architecture and design vulnerabilities are flaws built into how an organization might create an ICS. For example, Stouffer et al. (2015) describe a particular vulnerability where the ICS does not have a defined security perimeter. This lack of perimeter means the utility cannot ensure they have properly deployed the controls they rely on for protecting the confidentiality, availability, and integrity of the ICS. This can lead to potentially unauthorized access to the ICS and its data (information disclosure in the STRIDE process).

2.2.1 Configuration And Maintenance

Configuration and maintenance vulnerabilities occur when organizations like this one do not properly care for their ICS. Here, improperly configured ICS creates openings to the utility from unnecessary services, functions, ports, and protocols (Stouffer et al., 2015). Leaving default configurations from an installation will expose weaknesses and services to attack, leading to tampering or elevation of privileges under a STRIDE model.

2.2.1 Physical Access

Stouffer et al. (2015) cite allowing unauthorized personnel to have physical access to an ICS as a physical vulnerability the company must remediate. Failure to do so can lead to theft or destruction of data or hardware and unauthorized changes to the ICS through a variety of means. This is potentially a denial of service within STRIDE's framework.

2.2.1 Software Development

Software development vulnerabilities occur when developers use bad coding practices, such as improper data validation techniques on user inputs and received data to the ICS. Such a vulnerability will create many "vulnerabilities including buffer overflows, command injections, cross-site scripting, and path traversals" (Stouffer et al., 2015, p. C-9). Such attacks will lead to tampering or elevation of privileges under a STRIDE model.

2.2.1 Communications And Network

Communications and network vulnerabilities are those that happen when part of the ICS is communicating with each other. For example, if two ICS components communicate over a wireless network, Stouffer et al. (2015) note there needs to be sufficient data protection between the two points. These ICS elements transmit sensitive information on behalf of the utility and should have strong encryption in place to secure the communications link. A lack of encryption on these communications can lead to tampering noted in the STRIDE rubric.

2.3 Likelihood

Returning to NIST's (2012) *Guide for Conducting Risk Assessments*, the document defines likelihood as a "weighted risk factor based on an analysis of the probability that a threat can exploit a vulnerability (or set of vulnerabilities)" (p. 10). Unfortunately, determining the likelihood of a cyber attack is challenging. While the Cyber Risk Task Force of the American Academy of Actuaries (2021) notes such attacks are real in the modern world, using past data to determine future probability is hard because data sharing about past attacks is notoriously difficult to come by.

CISA (n.d.) within the US government requests the business community to share cyber threat data with them to better inform the larger cybersecurity community, but is not a requirement. Recently, the US Senate removed a requirement that would have mandated such reporting to CISA from the National Defense Authorization Act, dooming it to possibly passing at a later date (Starks, 2021). However, each critical infrastructure sector maintains an Information Sharing and Analysis Center (ISAC) and the utilities sector is no different.

The Electricity Information Sharing and Analysis Center (E-ISAC) (2020) serves to lessen the risk of cyber and physical security risks for its members; principally, the E-ISAC does this through its Cybersecurity Risk Information Sharing Program (CRISP) program. The company is presumably a member of E-ISAC based on its

geographical coverage and importance to the southern California region; E-ISAC does not publish its membership roster.

Irrespective of the determining actual probabilities that a threat actor could exploit a vulnerability, the utility needs only look at recent zero-day attacks announced to see there is always a high likelihood of attempted penetration. For example, researchers discovered a new vulnerability for Log4j (a key part of many Unix logging systems) and issued an alert on December 10, 2021 (National Institute of Standards and Technology, 2021). Four days later, attackers launched over 840,000 attacks on Log4j components (Jeffrey, 2021).

2.4 Impact

Impact is the amount of damage expected from a particular threat exploiting vulnerability for the company (National Institute of Standards and Technology, 2012). In more practical terms, impacts are usually costs associated with an ICS component failure and the need to replace it, paying to address reputational damage from angry customers, or for addressing new regulatory requirements on the heels of an attack it might have failed to address.

3. Calculating The Organization's Risk Exposure

NIST (2012) describes a method for document specific risks. The first step in the process is to determine the threat sources or events. Next, the company would determine the vulnerabilities it faces within its attack surface from each threat source. Third, the utility should determine the likelihood a threat would exploit each identified vulnerability and then determine what impact that exploitation would have on its operations. Figure 2 below outlines the overall risk assessment process.

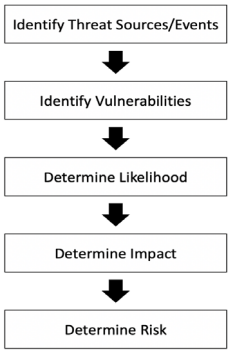


Figure 2: Risk Assessment Process

Note: Adapted from "Guide for Conducting Risk Assessments (SP 800-30, rev. 1)" by National Institute of Standards and Technology, 2012.

3.1 Threat-Vulnerability Pairs

A complete cybersecurity risk assessment for this company is beyond this paper. However, using the threats and vulnerabilities already presented, examples become obvious using the process described in Figure 2. For brevity's sake, this paper will only use the intentionally malicious threats described earlier: nation-state attackers, financially motivated criminals, and activists seeking to promote their cause. This paper randomly pairs each of these three threat sources with one of the five vulnerabilities discussed: lack of security perimeter, improper system configurations, unauthorized physical access to company facilities, improper data validation, and a lack of data protection between nodes. As noted previously, in a full risk assessment, the utility should map each threat source to each vulnerability.

This creates five threat-vulnerability pairs (P). Table 3 below provides them along with a brief narrative description of each. Each description also notes the associated MITRE (2020) attack technique in parentheses relevant to each P value.

Table 3: Organizational Threat-Vulnerability Pairs with Associated MITRE Attack Techniques

Identifier	Threat-Vulnerability Pair	Description
P1	Nation state - Perimeter	A nation state might attack an ICS device Edison International believes is protected by its perimeter but is not (*T0829*)
P2	Crime - System Configuration	A cyber criminal could attack an ICS with default system configurations (*T0885*)
P3	Activity - Physical	An activist group might gain access to an Edison electrical substation and destroy it (*T0879*)
P4	Nation state - Data Validation	A nation state could attack a database that stores ICS information using SQL injection techniques (*T0811*)
P5	Crime - Comms Protection	A cyber criminal might change vital communications from between ICS nodes if there is no encryption on the link (*T0830*)

Note: Adapted from "Guide for Conducting Risk Assessments (SP 800-30, rev. 1)" by National Institute of Standards and Technology, 2012 and "ICS Attack Techniques" by MITRE, 2020. Copyright 2020 by MITRE.

3.2 Risk Matrix

With qualitative values for likelihood and impact, the company can build a table that shows their intersection. The company should define these qualitative values (high, moderate, and low) in terms of their business needs. High likelihood might mean there is an openly available proof-of-concept piece of software for a known vulnerability on the internet, while a low impact could be that the affected system would not disrupt electrical generation and distribution. Table 4 below presents such a matrix combining likelihood and impact. This is only an example matrix; the company would need to tailor its values to its risk tolerance levels.

Table 4: Likelihood-impact matrix to show risk levels

		Impact		
		Low	Moderate	High
Likelihood	Low	Low	Moderate	Moderate
	Moderate	Moderate	Moderate	High
	High	Moderate	High	High

Note: Adapted from "Guide for Conducting Risk Assessments (SP 800-30, rev. 1)" by National Institute of Standards and Technology, 2012.

Table 4 implements a high-water mark process. Whatever the highest value is for either likelihood or impact becomes the risk level associated with that combination. The exception to this is the two combinations of low and high likelihood and impact. For those two items (*low likelihood, high impact* and *high likelihood, low impact*) this matrix splits the difference and marks both as moderate. For each threat-vulnerability pair in table 3, the company can enter the business appropriate likelihood and impact and then calculate the risks each pair presents to its business. Table 5 presents examples of what these results might look like using the matrix in Table 4.

Table 5: Likelihood, impact, and associated risk for each identified threat-vulnerability pair

Identifier	Likelihood	Impact	Associated Risk
P1	Low	High	Moderate
P2	High	High	High
P3	Low	Moiderate	Moderate
P4	High	High	High
P5	Low	Low	Low

Note: Adapted from "Guide for Conducting Risk Assessments (SP 800-30, rev. 1)" by National Institute of Standards and Technology, 2012.

Like most organizations, this utility does not possess unlimited resources and must prioritize where to get the greatest return on its investment in controls. Reviewing the information in Table 5, the company should focus its resources on resolving those concerns with the highest risks: TV2 (*a cyber criminal could attack an ICS with default system configurations*) and TV4 (*a nation state could attack a database that stores ICS information using SQL injection techniques).*

4. Mitigation Techniques for the Organization's High-Risk Pairs

How might the company address the two high-risk items? MITRE (2020) recommends specific countermeasures for each attack technique associated with the threat-vulnerability pairs.

4.1 P2 - Crime - System Configuration

To mitigate against ICS attack technique *T0885 - Commonly Used Port*, MITRE (2021b) recommends several changes. First, each controller deployed by the utility in the field should "require users to authenticate for all remote or local management sessions" (para. 3). MITRE's (2021b) second recommended mitigation is to disable all unnecessary ports and protocols; only allow what is absolutely necessary to function on the ICS component. The company can confirm this standard by using a commercially available Security Content Automation Protocol (SCAP) enabled vulnerability scanners to seek open ports and protocols. SCAP is "a suite of specifications for exchanging security automation content used to assess configuration compliance and to detect vulnerable versions of software" (National Institute of Standards and Technology, 2016, para. 1).

Third, it should install a network-based intrusion prevention system. Such a system analyzes specific information from the network attached ICS components to seek the characteristics that show an attack is occurring and prevent the attacking traffic from reaching its target (Wang & Xu, 2020). The last recommendation MITRE (2021b) makes is to segment the ICS components into their own distinct network space. Experts call this network segmentation. Specifically, the company should "[c]onfigure internal and external firewalls to block traffic using common ports that associate to network protocols that may be unnecessary for that particular network segment" (para. 3).

4.2 P4 - Nation state - Data Validation

To address the concerns in *T0811 - Data from Information Repositorie*s, MITRE (2021a) similarly makes a series of recommendations. First, the company must encrypt the information in its databases. There are many options for such an encryption scheme; the company should find the one that provides the correct balance between key management and flexibility (Ocenas et al., 2020). Second, the company should make use of a Privileged Account Management (PAM) solution that will "minimize permissions and access for service accounts to limit the information that may be exposed or collected by malicious users or software" (MITRE, 2021a, para.1). There are several available for purchase and implementation.

Third, MITRE (2021a) recommends the company restrict permissions to files and directories on the database server. This will prevent threat actors from interacting with and collecting data from the database server. This coincides with the fourth recommendation: ensuring users and groups have only the permissions via a properly configured Identity and Access Management (IdAM) program (MITRE, 2021a). Finally, the company should "develop an auditing mechanism to conduct periodic reviews of accounts and privileges for critical and sensitive repositories" (para. 3)

4.3 Beyond the High-Risk Findings

Once the utility addresses these two high-risk threat-vulnerabilities pairs, they can then focus on the moderate-risk items (P1 and P3) before moving on to the low-risk item (P5) presented in this example risk assessment. For each of the techniques listed in Table 3 and the calculated risks in Table 5, MITRE (2020) provides specific mitigation tactics it should implement.

5. Conclusion

This utility provides energy for an important part of the world. Understanding the threats and vulnerabilities companies face will improve cybersecurity across the country's critical infrastructure. Within the utility sector, risks to Industrial Control Systems (ICS) remain high. Cybersecurity and Infrastructure Security Agency (CISA) (2021) has released detailed advisories surrounding ICS.

Tried-and-true processes exist for assessing risks. While there are many processes for conducting a risk assessment, the US government's process is straight-forward. This particular utility company should find the process that best meets their business needs. In the process described previously, the company must identify its threats and understand its vulnerabilities. From there, they should estimate, with key management officials, the likelihood of those threat-vulnerability pairs affecting operations. The company can rank each threat-vulnerability pair, and its associated likelihood and impact, using a standard risk matrix. This will allow it to focus their resources on those pairs most likely to cause significant damage to its operations and business.

References

Abrams, M. & Weiss, J., 2008. Malicious control system cyber security attack case study–Maroochy Water Services, Australia.

Cart, J., 2021. California's 2020 Fire Siege: Wildfires by the numbers. CalMatters. Available at: https://calmatters.org/environment/2021/07/california-fires-2020/ [Accessed January 15, 2022].

Census Bureau, n.d.. U.S. Census Bureau quickfacts: Los Angeles County. QuickFacts. Available at: https://www.census.gov/quickfacts/fact/table/losangelescountycalifornia,CA/PST045219 [Accessed December 10, 2021].

Cherepanov, A., 2017. WIN32/Industroyer: a new threat for industrial control systems.

Cyber Risk Task Force, Casualty Practice Council, 2021. Cyber risk toolkit.

Cybersecurity and Infrastructure Security Agency, n.d.. Systemic cyber risk reduction. Cybersecurity and Infrastructure Security Agency. Available at: https://www.cisa.gov/systemic-cyber-risk-reduction [Accessed January 9, 2022].

Electricity Information Sharing and Analysis Center, 2020. E-ISAC long-term strategic plan update.

Fontinelle, A., 2021. Getting a grip on holding companies. Investopedia. Available at: https://www.investopedia.com/terms/h/holdingcompany.asp [Accessed December 10, 2021].

Hughes, R.A., 2020. If California were a country. Bull Oak Capital. Available at: https://bulloakcapital.com/blog/if-california-were-a-country/ [Accessed December 10, 2021].

Jeffrey, C., 2021. LOG4J flaw turns into pandemic with over 840,000 attacks initiated within 72 hours. TechSpot. Available at: https://www.techspot.com/news/92633-hackers-launch-over-840000-attacks-through-log4j-flaw.html [Accessed January 9, 2022].

Kaspersky, 2021. What's the difference between a virus and a worm? www.kaspersky.com. Available at: https://www.kaspersky.com/resource-center/threats/computer-viruses-vs-worms [Accessed December 16, 2021].

Khodabakhsh, A. et al., 2020. Cyber-risk identification for a digital substation. Proceedings of the 15th International Conference on Availability, Reliability and Security.

LinkedIn, 2021. Adam Tuzzolino. LinkedIn.com. Available at: https://www.linkedin.com/in/adam-tuzzolino-40a7596 [Accessed December 10, 2021].

MITRE, 2020. ICS attack techniques. Techniques - Attacks ICS. Available at: https://collaborate.mitre.org/attackics/index.php/All_Techniques [Accessed January 8, 2022].

MITRE, 2021. T0811 - data from information repositories. Data from Information Repositories - attack ICS. Available at: https://collaborate.mitre.org/attackics/index.php/Technique/T0811 [Accessed January 9, 2022].

MITRE, 2021. T0885 - Commonly used port. Commonly Used Port - attack ICS. Available at: https://collaborate.mitre.org/attackics/index.php/Technique/T0885 [Accessed January 9, 2022].

National Institute of Standards and Technology, 2012. Gaithersburg, MD: US Department of Commerce.

National Institute of Standards and Technology, 2016. Security content automation protocol: CSRC. CSRC. Available at: https://csrc.nist.gov/Projects/Security-Content-Automation-Protocol/faqs [Accessed October 14, 2021].

National Institute of Standards and Technology, 2021. CVE-2021-44228. NVD. Available at: https://nvd.nist.gov/vuln/detail/CVE-2021-44228 [Accessed January 9, 2022].

National Institute of Standards and Technology, n.d.. Baseline configuration - glossary. CSRC. Available at: https://csrc.nist.gov/glossary/term/baseline_configuration [Accessed November 24, 2021].

Ocenas, M. et al., 2020. Security and encryption at modern databases. Proceedings of the 2020 4th International Conference on Cryptography, Security and Privacy, pp.19–23.

Office of the Director of National Intelligence, 2021. Washington, DC: US Government.

Ovaz Akpinar, K. & Ozcelik, I., 2018. Development of the ECAT preprocessor with the Trust Communication Approach. Security and Communication Networks, 2018, pp.1–16.

OWASP, 2021. Attack surface analysis cheat sheet. Attack Surface Analysis - OWASP Cheat Sheet Series. Available at: https://cheatsheetseries.owasp.org/cheatsheets/Attack_Surface_Analysis_Cheat_Sheet.html [Accessed October 14, 2021].

Radanliev, P. et al., 2018. Future developments in cyber risk assessment for the internet of things. Computers in Industry, 102, pp.14–22.

Shostack, A., 2014. Threat modeling: designing for security, Wiley.

Spenneberg, R., Bruggemann, M. & Schwartke, H., 2016. PLC-Blaster: a worm living solely in the PLC. Blackhat. Available at: https://www.blackhat.com/docs/asia-16/materials/asia-16-Spenneberg-PLC-Blaster-A-Worm-Living-Solely-In-The-PLC-wp.pdf [Accessed January 3, 2022].

Starks, T., 2021. Cyber incident reporting mandates suffer another Congressional setback. CyberScoop. Available at: https://www.cyberscoop.com/cyber-incident-reporting-ransomware-payments-congress-ndaa/ [Accessed January 9, 2022].

Stouffer, K. et al., 2015. Guide to industrial control systems (ICS) security. NIST Special Publication 800-82, pp.1–247.

Wang, D. & Xu, G., 2020. Research on the detection of network intrusion prevention with SVM based optimization algorithm. Informatica, 44(2).

Webb, E.L., 2016. The internet of things: cybersecurity, insurance, and the national power grid. Natural Resources & Environment, 30(4), pp.35–39.

Xiong, W. & Lagerström, R., 2019. Threat modeling – A systematic literature review. Computers & Security, 84, pp.53–69.

On the Use and Strategic Implications of Cyber Ranges in Military Contexts: A Dual Typology

Andrew C Dwyer[1], Kathrin Moog[2], Jantje Silomon[2] and Mischa Hansel[2]
[1] Information Security Group, Royal Holloway, University of London, Egham, UK.
[2] Institute for Peace Research and Security Policy at the University of Hamburg, Germany.

andrew.dwyer@rhul.ac.uk
moog@ifsh.de
silomon@ifsh.de
hansel@ifsh.de

Abstract: The use of simulated environments in cybersecurity – cyber ranges (CRs) – has become a popular method and tool to support training and education, assess system vulnerabilities, as well as to test and probe computer networks. Yet, CRs can be adopted to improve and enhance adversarial skill sets, tools, and operations that are attractive for military applications. This paper develops a strategic typology on how CRs have been used and adopted in military contexts (CRiMCs), where states have turned to CRs as one method to build cyber capacity, address proportionality and responsibility of cyber operations, as well as offer training and education. We identify CRiMCs that offer two strategic purposes: 1) reserved for sovereign use and capability development and 2) those intended to support cyber capacity building through domestic resilience and collaborative inter-state exercises. We thus ask, why do states establish sovereign cyber ranges 'on top' of being involved in collaborative ones? Why and how do they differ? To answer such questions, this paper delves into both the crucial technical components that support each CRiMC type and their implications by offering exemplars from five states (Lithuania, Norway, Slovenia, the Netherlands, and the USA). The paper concludes with some preliminary thoughts on future research avenues on CRiMCs and their implications for the use and governance of state cyber capabilities.

Key Words: Cyber Ranges; Cyber Operations; Cyber Capacity Building; Typology; Cyber Strategy

1. Introduction

Establishing environments for testing and training are core to military thinking and development. Within cybersecurity, as computational networks have spread across society, this has stimulated a burgeoning industry offering training and assessing system vulnerabilities through the provision of virtual environments – cyber ranges (CRs). These have become popular in commercial and research applications, from testing cybersecurity response readiness to education and training, as much as they have had a sustained history and development in their use by militaries (Davis and Magrath, 2013). However, there has been relatively little examination of the *strategic* use of CRs in military contexts (CRiMCs) beyond an analysis of their technology and little on how they have been used by states to enhance their cyber capabilities, operations, and campaigns.

The building of cyber capabilities and their use to engage in inter-state competition has become increasingly commonplace, with a growing diffusion of military cyber forces across the world dedicated to such activity (Blessing, 2021; Smeets, 2022a). As cyber capabilities have stretched beyond 'cyber powers' such as the United States, Russia, and China, a greater number of states are seeking to exploit computational networks for both pre-emptive defence and to pursue various effects against adversaries. Such activity has frequently occurred beyond the conventional bounds of war, in what Kello (2017) identifies as a condition of 'unpeace'. Amid a strategic posture of persistent engagement by states through continuous operations and campaigns rather than standalone vulnerabilities or 'weapons', as argued in contemporary US doctrine (Fischerkeller et al., 2022), an increased operational tempo requiring regular assessments and investment to ensure operational effectiveness is required. The proliferation and use of state cyber capabilities has likewise been accompanied by a greater attention to the applicability of international law in cyberspace and commitments to non-binding rules on responsible behaviour (Painter, 2021). Thus, there is an ethical as well as normative commitment to ensure cyber operations are precise, proportionate, and responsible in their use.

We examine the application of CRs in this broader military context, identifying how they could be strategically used in a changing operational landscape. Although CRiMCs have existed since at least the early 2010s, such as the US National Cyber Range (US NCR) initiated by DARPA (Davis and Magrath, 2013; Rosenstein and Corvese, 2012). This does not only apply to international military cooperation (Miller, 2022), but also their domestic economic opportunities as well as sovereign capabilities as part of NATO's Cyber Range developed through Foundation CR14 in Estonia (2022). This paper then provides two key contributions. First, it develops a strategic typology of CR application to understand their integration and use by states in military contexts. These are split

into two categories of CRiMCs, centred on pursuing sovereign objectives and capabilities on the one hand, and demonstrating and enabling cyber capability building measures – both domestically and internationally – on the other. Second, it explores this typology to understand the strategic implications of CRiMCs for states. The paper then proceeds by: 1) offering a background on CRs and their use in military contexts; 2) setting out our empirical analysis and giving a brief overview to five states and developing a typology of CRiMCs into those for sovereign use and capacity building; before, 3) discussing the strategic implication of CRiMCs for states and their future applications.

2. Cyber Range Architecture and Use

CRs have become popular across industry and strategically for state actors, for example within China (Cary, 2022). Defined as "interactive, simulated platforms and representations of networks, systems, tools, and applications" (National Initiative for Cybersecurity Education, 2020, p. 3), they can be used for a variety of purposes, but primarily for training and skills development (Turčaník, 2020). This has led to a range of typologies on their architecture and use. For instance, academic literature makes a distinction between CRs and testbeds (Davis and Magrath, 2013; Ukwandu et al., 2020), with the latter often referring to the testing of operational technologies (OT). Much work has sought to likewise create different categories of use to apply to different types of groups and their technical architecture (Karjalainen and Kokkonen, 2020; Urias et al., 2018), as well as their military use (Debatty and Mees, 2019). Päijänen et al. (2021), for instance, identify both 'government' and 'military cyber defence capabilities' as target groups for CRs, and identify international exercises by states as one implementation. However, apart from their application to capacity building, their utility for the deployment of cyber operations and capabilities by states is underexplored. We therefore bridge an analysis of strategic use by states for their cyber capabilities and their technical manifestation, as we explore in our typology below.

CRs can, to some extent, be seen as a compromise: a large-scale and flexible environment allowing for complex scenarios that are predominantly virtualised (but can include emulation and hardware integration). CYBERWISER (2020), for example, uses a three-tiered categorisation based on functional capabilities:

- a simple, pre-defined, network-accessible but limited environment;
- a locally accessible infrastructure into which malware cannot be introduced; and,
- a large, complex, locally accessible infrastructure where all equipment and devices are provided by the cyber range vendor/operator, and malware may be safely run.

Furthermore, CRs cover a vast variety of systems and thus technical arrangements – a range focused on an industrial control system (ICS) will have limited use cases compared to generic 'off-the-shelf' set-ups more common in commercial and research applications. Yet, certain basic elements will be recognisable, including the three key architectural components: underlying infrastructure, orchestration, and range capabilities (ECSO, 2020; Ukwandu et al., 2020).

2.1 Three Architectural Components

Historically, CRs underlying *infrastructure* tended to rest entirely on dedicated hardware, such as utilising large, localised server racks, with systems built 'directly on top' of physical infrastructure. While such dedicated systems can remain important for some CRiMCs – both for enhanced security and enabling high fidelity capacities, particularly when simulating highly specialised or proprietary software and hardware, they continue to be expensive to construct and often lack flexibility. Coupled with advances in virtualisation and cloud technologies, CRs can be adapted to be economically efficient and iteratively updateable, which have gained widespread acceptance, particularly for commercial and research focussed CRs.

Orchestration combines two main aspects, the CR's environmental set-up and its management. The former includes, for example, automated configuration and coordination that becomes important when CRs are required to be reiterated frequently. Whereas the latter refers to elements such as attack scheduling and event injection to enable realistic scenarios.

Range capabilities are split into two main parts. The first is a type of learning management system used by the white team and can include scoring, guiding, and/or decisions. It can also be used to inform a green team, whose purpose is to maintain and repair/fix issues. The second is the target system, including their underlying

infrastructure (e.g., endpoints, servers, storage, applications, and networks), security operations (e.g., SIEMs and IDS), and the 'outside' (e.g., wider domains, proxies, DNS). While these will be present in both CRs and CRiMCs, and blue/red teaming might be conducted, the detail and complexity will vary. Similarly, so will the system response, which could be scripted, automated, live, or a combination thereof.

2.2 Adversarial Uses

Many scenarios might test how well an organisation responds against adversary behaviours and malware, such as well-known examples like WannaCry or Emotet. Or in the case of more OT-centred ranges, malware such as Triton, first discovered at a Saudi Arabian petrochemical plant in 2017, could be used. However, they can be deployed for adversarial uses too, which is critical for some CRiMCs applications. They may be used to build simulated environments of adversaries which states can test their capabilities against. A highly advanced and now infamous example – which may be described as a CRiMC – are those created for the Olympic Games broader suite of operations, commonly abridged to the Stuxnet malware (Zetter, 2014). In preparation for a series of interconnected operations against the Natanz nuclear enrichment facility in Iran by the USA, Israel, and other states, extensive CRiMCs were constructed. This hardware-software cyber testing was at least partially conducted at the Dimona nuclear plant in southern Israel as well as in the US at Oak Ridge (ibid). Thus, CRiMCs may not be singular or concentrated in one place, but constructed of various components, with shared data and facilities, or even testing for different components. Although Olympic Games and Stuxnet may represent the one – and only – publicly known example of the extensive fine-tuning of cyber capabilities through testing, the result was a stealthy and, importantly, highly-targeted, and discriminate operation that did not cause significant impact outside of its intended target. Despite its wide-spread propagation, it was designed to lay dormant before deletion, increasing its covert appeal and reducing the chances of its detection and consequent unintended impacts.

Despite the reported in-depth capabilities on display for Stuxnet, most states and cannot financially sustain or deploy the required skills for such testing alone. They may also have military cyber forces that can only engage in cases of armed conflict, i.e., only 'above' the threshold. Therefore, many states that may wish to test their capabilities are reliant on either international cooperation or on private enterprise, primarily from the defence and aerospace industry, to develop, maintain, and implement CRiMCs. These can fulfil a training role as much as other non-military and non-offensive capabilities testing. In most cases, generic set-ups for adversarial purposes are unlikely to be used unless states wish to test out how susceptible a common environment may be. Detailed target knowledge – sustained by significant intelligence collection – will not only improve operational effectiveness but could also help prevent collateral damage or unintended consequences, a requirement for a 'responsible' actor. Yet detailed knowledge of the target's environment, particularly its technical components, requires insights that may only be acquired by already having presence inside an adversary's environment. Knowing that one specific PLC is used without its detailed integration and implementation is unlikely to be sufficient. Instead, various component and subcomponent vendors might need to be known, or details such as the frequency band of motor spin systems, as Stuxnet demonstrated.

3. CRiMC Typology

CRs thus exhibit a wide range of applications, uses, and varying technological requirements, leading to greatly differing definitions and their application in military contexts. Using the categories laid out by CYBERWISER (2020), for example, only the third– a large, complex, and locally accessible infrastructure - could be considered relevant. However, underlying *infrastructure* options would have to include cloud-based components, at least in some cases. This aspect, as well as the range capabilities, are likely to diverge the most between CRs and CRiMCs, though again dependent on their intended use. In this section, we develop a typology of CRiMCs from research of public documentation: first, those reserved for sovereign capabilities by states and second those used to enable domestic and international cyber capacity building.

3.1 Methodology

To identify CRiMCs, we analysed public documentation primarily available in English for forty-six states in the European region, Canada, and the USA up to August 2022. This included the analysis of national cyber strategies, official announcements, as well as public news sources to understand the diffusion, and use, of CRiMCs by states.

Our preliminary analysis then identified thirty-six states using CRiMCs, demonstrating their often-complex arrangements. For example, Spain's Joint Cyber-Defence Command used the Minsait Cyber Range system by the private enterprise Indra to conduct both capacity building, information sharing, as much as to "conduct and execute military operations in cyberspace" (Indra, 2019). The quote is however demonstrative of the openness to interpretation of such materials – and does not signify, for example, that offensive cyber operations are tested on this CRiMC.

The publicly available information on CRiMCs is scarce, resulting in several limitations. When there is information, it is often brief and lacks details to appropriately assess how a CRiMC is used by a particular state. In our research, there is little, if any, technical detail on CRiMCs, their funding or intended use. With a few exceptions, such as the US NCR with more regular updates, many sources are vague, difficult to corroborate, and timely information is not routinely maintained. This is also compounded by the limitation of reviewing predominantly English-language documentation, which is likely to restrict insight into documentation produced by various states.

3.2 CRiMC Analysis

Table 1 presents our initial analysis of data from forty-six states as of August 2022, demonstrating a wide variety of states that engage in the usage of CRiMCs, with many owning CRs to use for capacity building, particularly domestically. However, far fewer have publicly declared CRiMCs for sovereign use that are unambiguous. The analysis of public documentation resulted in either a confirmation of existence or use ('✓'), sufficient to deem that it does not exist or is used ('✗'), or for situations of ambiguity or lack of information ('?').

Therefore, the data presented offers only an indicative analysis of how states are primarily organising their CR capabilities, where their sovereign use and ownership emerged through the data analysis. This typology then emerged during the analysis of states' strategic use of CRs between sovereign and capacity building capabilities that we define below as:

1. Sovereign CRiMCs: CRs for the exclusive development of sovereign cyber capabilities; primarily for honing active cyber operations and campaigns against adversaries but may expose vulnerabilities that can be used for improving a state's domestic cyber security.
2. Capacity Building CRiMCs: CRs to enhance defensive capability, complement and federate CRs, as well as share information and training both domestically and between states.

Below we present five cases in greater detail to demonstrate the different uses and ownership of CRiMCs before detailing how they technically and strategically differ in 3.3. Each case has a short description of its regional position and how it uses CRiMCs. This is not intended to be a systemic representation of CRiMCs but rather a descriptive interpretation of the range of configurations for their use and ownership. This means that we seek to include states that engage in capabilities across both sovereign and capacity building use, engage internationally and domestically, and how different models of ownership ranging from being exclusively in miliary settings to more collaborative set-ups.

3.2.1 Lithuania

A Baltic state using a cyber capacity CRiMC based within national defence including international cooperation.

In 2022, *Virtualus Kibernetinis Poligonas* was deployed as a new CRiMC (Lithuanian Ministry of National Defence, 2022), with an investment of €800,000. It has a range of uses, enabling the "simulation of cyber-attacks of different levels in a safe closed environment" (ibid). This includes CTFs as well as real-time team exercises, accessed remotely. It is intended to be used by staff from defence, cybersecurity actors in Lithuania, as well as foreign partners. The CRiMC is run by the National Cyber Security Centre under the Ministry of National Defence.

3.2.2 Norway

A Nordic state using a cyber capacity CRiMC in an academic setting, with defence funding, as well as with international cooperation.

The Norwegian Cyber Range is based within the Norwegian University of Science and Technology (NTNU) to enable cooperation between government, business, and academia launched in 2018 (NTNU, 2022), funded with ~€7 million. The CRiMC has both remote and physical access (from 2022) with support from the Norwegian

armed forces. It is also part of an "Open Cyber Range" with Estonia intended for cybersecurity training in the private and education sectors (Norwegian Ministries, 2019).

Table 1: An overview of CRiMC use and ownership according to our outlined typology.

State	CRiMC use in public documentation	CRiMC owned for sovereign use	CRiMC owned for cyber capacity building
Albania	✓	×	×
Armenia	?	?	?
Austria	✓	×	✓
Azerbaijan	?	?	?
Belarus	?	?	?
Belgium	✓	×	✓
Bosnia and Herzegovina	×	×	×
Bulgaria	✓	×	✓
Canada	✓	×	✓
Croatia	✓	×	×
Cyprus	×	×	×
Czechia	✓	×	✓
Denmark	✓	?	?
Estonia	✓	✓	✓
Finland	✓	×	✓
France	✓	×	✓
Georgia	✓	×	×
Germany	✓	×	✓
Greece	✓	×	✓
Hungary	✓	×	✓
Iceland	?	×	×
Ireland	✓	×	✓
Italy	✓	?	✓
Kosovo	×	×	×
Latvia	✓	×	✓
Lithuania	✓	×	✓
Luxembourg	✓	×	✓
Malta	✓	×	✓
Moldova	?	?	?
Montenegro	?	?	?
The Netherlands	✓	✓	✓
North Macedonia	?	?	?
Norway	✓	×	✓
Poland	✓	×	✓

Portugal	✓	×	✓
Romania	✓	×	✓
Russia	✓	?	✓
Serbia	✓	×	✓
Slovakia	✓	×	✓
Slovenia	✓	×	✓
Spain	✓	?	✓
Sweden	✓	×	✓
Switzerland	✓	×	✓
Turkey	?	?	?
Ukraine	✓	×	✓
United Kingdom	✓	?	✓
USA	✓	✓	✓

3.2.3 Slovenia

A central European state with a cyber capacity CRiMC supported by another state.

Slovenia's CRiMC is supported by €12.7 million from the United States' European Command to incorporate a range of different activities and actors from academia, private research, and civil society (U.S. Embassy in Slovenia, 2020). This is being combined in the Slovenian Ministry of Defence with its Security Operations Centre (Slovenian Republic, 2022). There is no detail on its technical set-up.

3.2.4 The Netherlands

A western European state with a CRiMC for both sovereign and cyber capacity capabilities.

Limited information about a sovereign cyber range built for Defence Cyber Command (DCC) in 2016 by private defence contractor Thales was identified. It is used for training and development as well as "to acquire and test new cyber-defence techniques at an early stage." As the then-Commander, Hans Folmer, said, "[i]t is a facility at which many forms of cyber operations can be simulated" (Thales, 2016). There is little information post-2016 about subsequent activities beyond the initial press release, which stated it would only last for 3 years.

3.2.5 United States

A North American state with multiple CRiMCs for both sovereign and cyber capacity capabilities with international cooperation.

The US National Cyber Range (NCR) is one of several CRiMCs, with the NCR initiated by DARPA to build an Internet-like environment, since at least 2008 (DARPA, No Date), with a range of different technical components that enable both classified and unclassified use by a variety of actors. In 2021, the U.S. Army awarded part of a $2.4bn contract for a range complex to BAE Systems (2021). Similarly, there has been clear offensive capability on the range, as much as it is used for multi-country exercises, such as annual USCYBERCOM exercises, with physical and remote, cloud-based access. As a recent Department of Defense press release noted, "the NCR enables DOD to conduct virtual, combined and joint cyberspace training, exercises, mission rehearsals, experiments and certifications… enabl[ing] a high degree of collaboration, development of U.S. and allied cyber tactics, techniques and procedures for defensive cyber missions" (U.S. Cyber Command, 2021). This is only a limited snapshot of the US NCR, which has developed and evolved over time, with a great deal of detail on its composition and activities.

3.3 Sovereign and Capability Building Uses

3.3.1 Sovereign CRiMCs

The identification of sovereign uses of CRiMCs is difficult but is ambiguously implied across various documentation and most explicitly in the case of the Thales-built cyber range in the Netherlands. Likewise, sovereign use is enabled through the segregation between classified and unclassified networks in the US NCR. Such sovereign uses will have different technical features compared to those for capacity building due to their alternative strategic nature. They are more likely to:

- Require emulation or high-fidelity simulations of environments in which to test capabilities. This means they resemble arrangements closer to testbeds with cyber-physical capacities to identify precise analysis for cyber operations to enhance responsibility and promote covertness.
- Embed hardware when wishing to robustly test against specific assets and adversaries.
- Require flexible and effective environments and scenario planning, with integration of active intelligence for unique implementations.
- Have greater attention to authentication, vetting of personnel, and segregation of networks (to limit attacks against CRiMCs themselves).

This is not an exhaustive list and the technological requirements for sovereign CRiMCs will require alternative arrangements, which, in turn, are discussed less in the public domain. Hence, it is difficult to appropriately test empirically these claims from public documentation.

3.3.2 Capacity Building CRiMCs

Capacity building – particularly through training and education – is a key feature of most CRs, whether that is in corporations, universities, or in enhancing the cybersecurity of state bureaucracies. This is replicated in CRiMCs, where it is possible to sub-divide their use according to:

1. *Domestic Capacity Building*: This is seen in almost all CRiMCs, where their dual – and even multiple – uses extend in complex ways. For example, with the US NCR, the invitation of improving domestic civilian cybersecurity is deemed crucial. This is even more so in the case of the Norwegian cyber range, which is primarily focused on such capacity building albeit supported by the armed services.
2. *Inter-State Capacity Building*: This is where states are sharing CR resources – such as through wargaming – to enhance another state's (and potentially their own) cybersecurity. Lithuania seeks to invite foreign stakeholders into their CRiMC, whilst the US explicitly uses its NCR to build allied capacities for security. In the most extreme case, we could see the US's funding of Slovenia's CRiMC as evidence of this form of capacity building indirectly.

In contrast to sovereign CRiMCs, Capacity Building CRiMCs are about demonstrating shared commitments to improve cybersecurity (and in the process of wargames scenarios, *abstracted* counter-adversary capabilities). Such capacity-building CRiMCs share similar technological attributes – with focuses on CTFs, integration of third-party actors often via cloud technologies, can be similar to those built by commercial providers, and have environments that are more generically modifiable with known vectors. Unlike most CTFs, however, these CRiMCs not only have dedicated infrastructure and resources, but are also more formalised in state or state-sponsored institutions.

Capacity building can also include the explicit sharing of training, information on best practice and the pooling of cyber ranges to extend capacity. One example of this is through the EU's Cyber Federations project, launched in 2017 and renewed in 2021 (European Council, 2021), which seeks to interconnect several EU cyber ranges into a larger cluster or CRiMCs developed at NATO. This is not too dissimilar from the US NCR, which combines several different technical components to build a larger cyber range with shared facilities that can be remotely accessed.

3.4 Making a Distinction Between CRiMCs

In the technical distinction of CRiMCs between sovereign and capacity building purposes, we do not assert that they must be mutually exclusive. CR architecture may permit both sovereign and capacity building (for example, as the US NCR does). Rather, attention must be paid to the CR orchestration and range capabilities in context. There is strong technological and strategic reasoning – to maintain secrecy, enable covert activity, and reducing the attack surface of CRiMCs – for separating out sovereign and capacity building capabilities. Assuring and

protecting CRs when engaging active sovereign capabilities – such as cyber operations against adversaries – requires greater investment. It is unlikely that most states would be able to have the expertise and capacity to configure both uses on the same CR, even when federated, apart from a very limited number of actors, such as in the US. There is therefore a close connection between the strategic ambitions for the use of CRs and their technical manifestation, as much as the latter then shapes the strategic possibility for most states due to the cost and expertise required to build and maintain sovereign CRiMCs. This is likely why there are many states engaging in CRiMCs, particularly with capacity building both domestically and internationally. This is because they frequently replicate 'off-the-shelf' commercial products, with preconfigured information on such CRs being transferable, as much as they offer more advantageous security and economic benefits for states. However, the public documentation on sovereign use is likely reflective of the strategic ambitions of states to engage in sustained cyber operations – such as the United States and the Netherlands. Other states may not wish to publicly declare CRs for sovereign capabilities to avoid declaring a more active strategic posture, thus this analysis underreports the extent of their use by states.

4. Strategic Implications

Based on the CRiMCs typology and analysis outlined above, there are multiple strategic implications for states and the practice of cybersecurity. This section offers three pertinent, but not exhaustive, set of issues: first on the maintenance of sovereign capabilities, the development of responsible behaviour and second on what the use of CRiMCs means for state signalling. More detailed exploration of deterrence aspects and potential escalation risks is not within the scope of this paper but subject of further research.

4.1 Maintaining Sovereign Capabilities

As recently pointed out by Smeets (2022b), states are reluctant to share cyber capabilities with allies and partners, partly due to the transitory nature of these capacities. There is an increased risk of detection that could 'burn' an information advantage, for instance with backdoors or zero-days. Such thinking can be applied to CRiMCs, as sharing technical details about the CR's architecture, orchestration or capabilities may expose details on a state's intelligence collection and integration that could be maliciously used to reverse engineer the CR or introduce false information and data. Indeed, a CRiMC for sovereign capabilities may assist in greater precision, proportionality, and operational tempo that states may see as increasing the effectiveness of cyber operations. Instead of the risk of transferring sovereign capabilities and losing a potential competitive advantage, states therefore engage in less risky capacity-building measures including cyber exercises, intelligence sharing, or funding CRiMCs as the US has in Slovenia even if they have a sovereign CRiMC. This supports the assertion that states wish to maintain exclusive technical competency at some level, even as joint exercises, such as NATO's Locked Shields, are on the rise.

4.2 Responsible Behaviour

For years, cyber diplomats have discussed principles of due diligence and state responsibility in cyberspace. As part of the eleven non-binding norms adopted by various fora of the United Nations, states voluntarily agreed not to target critical national infrastructures nor target emergency computer response teams (United Nations, 2021). Beyond this, states have also noted that the targets of attacks must therefore not be indiscriminate and maintain proportionality. The use of CRiMCs, particularly those geared towards sovereign uses, could stimulate adherence to these norms through testing of both capabilities and increased training and education, especially for state militaries who only act during armed conflict and wish to maintain operational effectiveness. However, more empirical research is needed to assess whether CRiMCs are helpful in reducing collateral damage and in encouraging responsible and proportionate behaviour. There is also the question how to avoid capability gaps and operational friction between allies, particularly if only sovereign CRiMCs enable significant advancements in conducting discriminatory and proportional cyber operations.

4.3 Signalling

CRiMC's development and use could have important signalling effects. For example, they might partially substitute 'real weapons tests' by signalling advanced capabilities for training and operational planning. States sharing some access to CRiMCs of their allies could hope to benefit from an extended deterrence effect, such as

in the case of the US NCR. That said, some states do not have CRIMCs or do not advertise them, as seen in Table 1. Others barely acknowledge theirs, as for example in in the case of the Netherlands. Based on the signalling argument, the lack of a CRiMC or their explicit acknowledgement could perhaps be seen to undermine a state's cyber deterrence, however further research will be required to assess this. In general, we contend that certain cyber ranges, specifically those simulating specific target systems, are very unlikely to be revealed for signalling purposes, since doing so would dramatically reduce their operational value and come with a heightened risk of intelligence loss. Acknowledging the existence of cyber ranges that were used in preparation of past cyber operations could be a different matter though. However, states may refrain from this due to legal and political risks, i.e., the same reason why there is rarely self-attribution of specific cyber operations.

5. Conclusions

This paper has explored the role of states in cyber ranges for military contexts – CRiMCs. Through a preliminary analysis of forty-six states, and the use of five indicative cases of Lithuania, Norway, Slovenia, the Netherlands, and the USA, a strategic typology of their use by states was developed. They were divided into 1) sovereign CRiMCs for the development of sovereign cyber capabilities and 2) capacity building CRiMCs that enhance both domestic cybersecurity as well as foster inter-state use and information sharing. Through this, both the technical and strategic underpinnings for such a distinction were assessed. However, due to limited public documentation and ambiguous language, the paper does not claim to be able to assess the relative distribution of each type within the international system. Different methods of data collection, such as expert interviews, will be required to overcome the limitations of public documentation for technical detail as well as the strategic ambiguity that many states employ regarding cyber operations. Other questions relate to the strategic relevance of CRiMCs. Beyond improving the effectiveness and responsibility of cyber operations, they could play a role in signalling and thus have a deterring effect. Since the latter is dependent on a minimum degree of transparency, future studies could take a closer look into the differences of public communication about CRiMCs across states. Systemically applying CRiMCs in a capacity building manner could offer novel ways to share the burden and expense of CRs but could also entail risks, which will also be the focus of future research.

References

BAE Systems, 2021. BAE Systems selected for National Cyber Range Complex contract [WWW Document]. URL http://web.archive.org/web/20221029102424/https://www.baesystems.com/en/article/bae-systems-selected-for-national-cyber-range-complex-contract (accessed 10.29.22).

Blessing, J., 2021. The global spread of cyber forces, 2000–2018. Presented at the 2021 13th International Conference on Cyber Conflict (CyCon), IEEE, pp. 233–255.

Cary, D., 2022. Downrange: A Survey of China's Cyber Ranges (Issue Brief). Center for Security and Emerging Technology, Washington, D.C.

DARPA, No Date. The National Cyber Range: A National Testbed for Critical Security Research.

Davis, J., Magrath, S., 2013. A survey of cyber ranges and testbeds (No. DSTO-GD-0771). Cyber and Electronic Warfare Division, Defence Science and Technology Organisation, Department of Defence, Australian Government, Edinburgh, South Australia.

Debatty, T., Mees, W., 2019. Building a Cyber Range for training CyberDefense Situation Awareness, in: 2019 International Conference on Military Communications and Information Systems (ICMCIS). Presented at the 2019 International Conference on Military Communications and Information Systems (ICMCIS), pp. 1–6. https://doi.org/10.1109/ICMCIS.2019.8842802

European Council, 2021. EU defence cooperation: Council launches the 4th wave of new PESCO projects [WWW Document]. URL http://web.archive.org/web/20221029124750/https://www.consilium.europa.eu/en/press/press-releases/2021/11/16/eu-defence-cooperation-council-launches-the-4th-wave-of-new-pesco-projects/ (accessed 10.29.22).

Fischerkeller, M.P., Goldman, E.O., Harknett, R.J., 2022. Cyber Persistence Theory: Redefining National Security in Cyberspace. Oxford University Press.

Foundation CR14, 2022. CR14 NATO Cyber Range [WWW Document]. URL http://web.archive.org/web/20221031160920/https://cr14.ee/ (accessed 10.31.21).

Indra, 2019. The Joint Cyber-Defence Command completes its training with Indra Cyber Range [WWW Document]. Indra. URL

http://web.archive.org/web/20221024113226/https://www.indracompany.com/sites/default/files/190717_pr_indra_cyber_range_-_mando_conjunto_ciberdefensa.docx.pdf (accessed 10.24.22).

Karjalainen, M., Kokkonen, T., 2020. Comprehensive Cyber Arena; The Next Generation Cyber Range, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). Presented at the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 11–16. https://doi.org/10.1109/EuroSPW51379.2020.00011

Kello, L., 2017. The Virtual Weapon and International Order. Yale University Press, New Haven.

Lithuanian Ministry of National Defence, 2022. Cyber Range of the National Cyber Security Centre unveiled [WWW Document]. URL https://kam.lt/en/cyber-range-of-the-national-cyber-security-centre-unveiled/ (accessed 10.29.22).

Miller, M., 2022. NATO prepares for cyber war. Politico.

National Initiative for Cybersecurity Education, 2020. The Cyber Range: A Guide (Draft Guidance). NIST.

Norwegian Ministries, 2019. List of measures – National Cyber Security Strategy for Norway.

NTNU, 2022. Norwegian Cyber Range [WWW Document]. URL https://www.ntnu.no/ncr (accessed 10.29.22).

Päijänen, J., Saharinen, K., Salonen, J., Sipola, T., Vykopal, J., Kokkonen, T., 2021. Cyber Range: Preparing for Crisis or Something Just for Technical People? Presented at the ECCWS 2021 20th European Conference on Cyber Warfare and Security, Academic Conferences Inter Ltd, p. 322.

Painter, C., 2021. The United Nations' cyberstability processes: surprising progress but much left to do. J. Cyber Policy 6, 271–276. https://doi.org/10.1080/23738871.2021.2014920

Rosenstein, M., Corvese, F., 2012. A Secure Architecture for the Range-Level Command and Control System of a National Cyber Range Testbed, in: 5th Workshop on Cyber Security Experimentation and Test (CSET 12). USENIX Association, Bellevue, WA.

Slovenian Republic, 2022. Izvedbeni načrt za izpolnjevanje kriterijev in zavez Republike Slovenije za Stalno strukturno sodelovanje na področju varnosti in obrambe Evropske unije (PESCO).

Smeets, M., 2022a. No Shortcuts: Why States Struggle to Develop a Military Cyber-Force. C Hurst & Co Publishers Ltd, London.

Smeets, M., 2022b. Why NATO Countries Don't Share Cyber Weapons. Natl. Interest. URL http://web.archive.org/web/20221024160629/https://nationalinterest.org/blog/techland-when-great-power-competition-meets-digital-world/why-nato-countries-don%E2%80%99t-share-cyber (accessed 10.24.22).

Thales, 2016. Thales to build "Cyber Range", a cybersecurity training and testing facility for the Dutch Defence Cyber Command [WWW Document]. URL http://web.archive.org/web/20221029095814/https://www.thalesgroup.com/en/worldwide/press-release/thales-build-cyber-range-cybersecurity-training-and-testing-facility-dutch (accessed 10.29.22).

Turčaník, M., 2020. A Cyber Range for Armed Forces Education. Inf. Secur. 46, 304–310. https://doi.org/10.11610/isij.4622

Ukwandu, E., Farah, M.A., Hindy, H., Brosset, D., Kavallieros, D., Atkinson, R., Tachtatzis, C., Bures, M., Andonovic, I., Bellekens, X., 2020. A Review of Cyber-Ranges and Test-Beds: Current and Future Trends. Sensors 20. https://doi.org/10.3390/s20247148

United Nations, 2021. Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. United Nations.

Urias, V.E., Stout, W.M.S., Van Leeuwen, B., Lin, H., 2018. Cyber Range Infrastructure Limitations and Needs of Tomorrow: A Position Paper, in: 2018 International Carnahan Conference on Security Technology (ICCST). Presented at the 2018 International Carnahan Conference on Security Technology (ICCST), pp. 1–5. https://doi.org/10.1109/CCST.2018.8585460

U.S. Cyber Command, 2021. DOD's Largest Multinational Cyber Exercise Focuses on Collective Defense [WWW Document]. URL http://web.archive.org/web/20221029103330/https://www.defense.gov/News/News-Stories/Article/Article/2863303/dods-largest-multinational-cyber-exercise-focuses-on-collective-defense/ (accessed 10.29.22).

U.S. Embassy in Slovenia, 2020. U.S. Strengthens Collaboration with Slovenia on Cybersecurity with New Cyber Range [WWW Document]. URL http://web.archive.org/web/20221029084128/https://si.usembassy.gov/u-s-strengthens-collaboration-with-slovenia-on-cybersecurity-with-new-cyber-range/ (accessed 10.29.22).

Zetter, K., 2014. Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon. Crown Publishers.

Evaluation of Quantum Key Distribution for Secure Satellite-integrated IoT Networks

Andrew Edwards[1], Yee Wei Law[1], Ronald Mulinde[1], Jill Slay[1,2]
[1]UniSA STEM, University of South Australia, Mawson Lakes, Australia
[2]SmartSat Cooperative Research Centre, Adelaide, Australia

edwaj011@mymail.unisa.edu.au
yeewei.law@unisa.edu.au
ronald.mulinde@unisa.edu.au
jill.slay@unisa.edu.au

Abstract: There has been a dramatic increase in the number of Internet of Things (IoT) devices and their applications. Furthermore, there is a growing impetus to integrate IoT networks on a global scale, using satellites to expand the range of IoT connectivity into geographically remote areas. Ensuring the security of satellite backhaul for IoT networks is thus of paramount importance. The steady advance of quantum computing in recent years threatens to nullify classical cryptographic approaches based on assumptions of computational hardness, motivating the need for post-quantum cryptography. Quantum computing algorithms have been developed that, once a quantum computer of sufficient scale is realised, will be able to break classical cryptosystems efficiently (at polynomial-time complexity). A promising method of securing information against this threat at the physical layer has emerged in the form of quantum key distribution (QKD). QKD exploits the fundamental physical properties of light to guarantee information-theoretic security. Research into the application and standardisation of QKD to secure satellite backhaul, however, is still in its infancy. This paper presents a brief overview of the theoretical basis for QKD, whilst also providing a survey of contemporary QKD protocols. It evaluates the ability of these protocols to secure satellite backhaul in the context of a typical satellite-IoT network architecture. Furthermore, it highlights the vulnerabilities, as well as the technical challenges associated with this endeavour. Finally, it proposes directions for future research and development into protocols and standardisation for the satellite-integrated IoT domain. Several challenges must be overcome before QKD can evolve into a global-scale solution for securing satellite-IoT. Secret key generation rate remains very low in practical demonstrations of trusted-relay QKD satellite architectures. Further research is needed to overcome or mitigate the fundamental rate-distance trade-off before satellite QKD can be considered practicable in an IoT application. Alternatives that do not rely on trusted nodes are contingent on nascent technologies such as quantum repeaters and quantum memory. Whilst in theory QKD provides perfect information-theoretic security, it remains vulnerable to attacks that exploit imperfections in real-world equipment. Further effort is needed to develop QKD protocols that can safeguard against the aforementioned challenges.

Keywords: Internet of Things, satellite communications, quantum cryptography, quantum key distribution

1. Introduction

The Internet of Things (IoT) describes a family of technologies which enable physical "things" to sense and affect the world around them in an intelligent way, by communicating over the internet and processing information in the cloud. In this paper, we adopt the term *Machine-Type Device (MTD)* from Centenaro et al. (2021) to refer to a "thing" in the IoT network.

The number of MTDs connected to the Internet is rising rapidly. From 2016 to 2020, the total number of connected devices in the IoT rose from approximately 7 billion to more than 20 billion, and the number of connected devices per person increased from 0.8 to almost 3 over the same period (Cheng et al., 2017). This scale of networking presents a massive attack surface. There is a growing economic impetus to integrate IoT networks on a global scale using satellites to handle backhaul traffic (i.e., aggregated traffic between gateways in the IoT network) because the satellites can greatly expand the range of IoT services into remote or wide-spread areas, where wholly terrestrial networks are impracticable (Centenaro et al., 2021). Therefore, there is a strong motivation to find ways to guarantee the security of satellite-integrated IoT protocols well into the future.

Current techniques for securing IoT traffic rely on a public-key infrastructure and exploit the innate intractability of certain one-way (mathematical) functions, used in cryptography to ensure that unauthorised decryption is intractable to probabilistic polynomial-time adversaries (Diamanti et al., 2016). However, the rise of quantum computing threatens to undermine the efficacy of pre-quantum cryptographic schemes (i.e., cryptographic schemes designed without consideration of attacks using quantum computing). Shor's (1994) algorithm can provide superpolynomial speedup to classical algorithms for breaking (i) cryptosystems based on the RSA problem (e.g., RSA cryptosystems) and (ii) cryptosystems based on the discrete logarithm problem (e.g., elliptic-

curve cryptosystems, Diffie-Hellman key agreement); see (Bernstein and Lange, 2017). In particular, Gidney and Ekerå's (2021) implementation of Shor's algorithm can factor 2048-bit RSA integers in 8 hours using 20 million noisy qubits. Grover's (1996) algorithm reduces the effective size of symmetric keys by half. The existence of these algorithms proves that the rise of quantum computing poses a serious threat to traditional secure communication protocols, including those used by satellite-integrated IoT networks.

There is a great deal of research currently underway to develop solutions that mitigate the emerging threat of quantum computing to traditional data security paradigms. The research can be divided into two main streams:

1. **Post-quantum cryptography (PQC)**: Also known as quantum-resistant cryptography (QRC), this refers to the design and development of computationally secure cryptographic schemes against which quantum-computing adversaries have no intrinsic advantage over classical adversaries. In 2016, NIST initiated a process to develop and standardise quantum-resistant public-key cryptographic algorithms to augment existing public-key cryptographic standards. In 2022, after three review rounds, four algorithms were selected to be standardised and four other algorithms were shortlisted for further reviews (Alagic et al., 2022). One of the shortlisted candidates, namely Supersingular Isogeny Key Encapsulation (SIKE), was recently broken (Castryck and Decru, 2022), serving as a reminder of the challenges of basing cryptographic algorithms on hard computational problems.
2. **Quantum cryptography**: Unlike cryptographic schemes that rely on computational intractability to deter attackers, quantum cryptography makes use of the fundamental physical properties of quantum mechanics to achieve information-theoretic or unconditional security.

The need to develop PQC is urgent, and in some areas, it may already be too late to protect the confidentiality of data. The fields of quantum computing and PQC are in a horse-race to see which will be ready first. The parameters of the race are illustrated in Figure 1, where the start line is an arbitrary point in time before quantum computing and PQC reach maturity. The distance to the finish line is determined by the developments needed to make PQC a practical reality. Confidence in the security of a PQC algorithm takes time to grow even after the algorithm has been standardised, because it is proportional to the time the algorithm stays undefeated. PQC cryptosystems need to be integrated with existing standards/protocols if feasible, or otherwise new standards/protocols will need to be designed and developed (Bernstein et al., 2022). Despite their resource constraints, IoT devices will need to implement both pre-quantum and post-quantum cryptographic schemes (Fernández-Caramés, 2020). If a large-scale quantum computing system becomes practicable before then, global data security is at risk on an unprecedented scale.

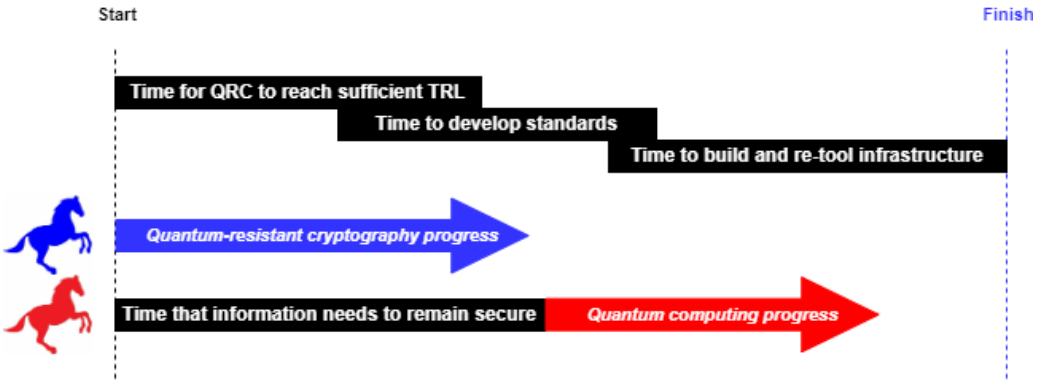


Figure 1: Illustration of the 'horse-race' metaphor between quantum computing and PQC progress.

The threat to global data security in fact more pressing than the preceding discussion would suggest, due to regulatory requirements on the length of time for which data needs to remain secure. Information that is transmitted over the IoT, even up to the day before PQC is fully in place, can still be harvested, stored and decrypted later once quantum computing reaches large-scale. For this analysis, "large-scale" refers to quantum computers of sufficient complexity to run Shor's algorithm. Quantum computing essentially has a head start equal to the maximum duration of time that electronic records need to remain secure. For example, for credit card information, this could be as little as one year, and for medical records it could be as much as 99 years (Commonwealth of Australia. *Archives Act 1983*).

The most widely studied and technologically mature sub-topic of quantum cryptography is quantum key distribution (QKD). QKD exploits the quantum mechanical properties of the Heisenberg uncertainty principle

and/or entanglement to encode a secret key using quantum states which are transmitted over a quantum channel that is theoretically immune to passive eavesdropping. These principles and their application to QKD will be described in Section 3.

QKD has been successfully applied to optical fibre networks for years, and QKD-enabled systems are offered as turn-key solutions for terrestrial networks. However, research into the application of QKD to wireless networks is still in its infancy, and application to satellite communications is lagging even further.

Nordholt et al.'s (2002) experiment first showed that QKD can be effectively utilised over a terrestrial free-space optical link. The experiment involved transmitting single photons with modulated polarisations from one place to another on the Los Alamos National Laboratory site. The researchers surmised that the technique could extend equally well to low-Earth orbit (LEO) distances.

More recently, there have been QKD experiments performed on actual space-ground optical downlinks from LEO, most notably using the satellite *Micius* under the auspices of the Chinese QUESS program (Liao et al., 2017) and *Tiangong-2 Space Lab* (Li et al., 2022). These experiments show that whilst their key generation rate is still too miniscule to be useful in an IoT backhaul application, the concept of satellite-based QKD is at least feasible and now needs only to be scaled up.

One of the main technical challenges limiting the deployment of QKD is the fundamental rate-distance trade-off. The path loss resulting from decoherence of entangled photons scales exponentially with distance, and hence the achievable secret key generation rate decreases exponentially with distance, even in free space (Takeoka et al., 2014). This limits the separation of two QKD-enabled communicating nodes to an order of hundreds of kilometres (Azuma et al., 2016). The use of a daisy chain of repeaters, in the traditional sense, to boost the range is untenable, due to the *no-cloning theorem*, a fundamental tenet of quantum mechanics which states that it is not possible to create a perfect copy of a quantum state (Wootters and Zurek, 1982). There is promising research into the development of a quantum version of the classical repeater, however such a repeater relies on the use of quantum memory for synchronisation purposes, which is still a nascent technology (Mastriani and Iyengar, 2020).

Only a handful of papers have considered the application of QKD to satellite-integrated IoT, and of those, most discuss it only in passing. Djordjevic (2020b) underscores the applicability of quantum methods to IoT, but does not specifically address the challenges of implementing quantum techniques to secure satellite-integrated IoT networks. It is apparent that whilst QKD is the most mature of the quantum cryptographic technologies under development, there is still a general lack of treatment in the literature on the viability and challenges in applying QKD specifically to the satellite-integrated IoT context. This paper seeks to provide a starting point for more focused research and development into this application area in the future.

2. Satellite IoT

There are two principal topologies for satellite-integrated IoT networks. In direct-to-satellite IoT, a satellite up-link is established directly from IoT device(s). In an indirect-access architecture, data from IoT devices is aggregated in the Network Layer using an LPWAN protocol such as LoRaWAN before being sent to a satellite for further routing (Fraire et al., 2019). At the destination, the data moves up the protocol stack to the Application and Business Layers (Zhou et al., 2021). Satellites are ideal for providing IoT connectivity in remote locations and/or over wide geographic areas where terrestrial communication technologies are impractical (Centenaro et al., 2021).

The four main security objectives for wireless networks (Zou et al., 2016) are equally applicable to satellite communications:

- **Confidentiality** - ensuring that unauthorised parties cannot decipher the communications between authorised parties involved in the communication session.
- **Integrity** - ensuring that all data transmitted by the sender is delivered intact, without modification, omission, or addition, to the recipient.
- **Authenticity** - ensuring that all parties involved in a communication session are who they claim to be.
- **Availability** - ensuring that a session can be set-up on demand and maintained for the required duration to enable communication to take place.

QKD can provide provable guarantees about the confidentiality and integrity of secret key exchange but does not address authenticity and availability requirements. However, there are choices that can be made in the architectural design of a QKD system to help ensure availability of the quantum channel, as we will discuss later.

3. Quantum Key Distribution

Quantum key distribution (QKD) is a method for generating and distributing symmetric cryptographic keys with information-theoretic security based on quantum information theory (ETSI, 2018). QKD typically uses a quantum channel in parallel with an authenticated classical channel to share a secret key between two parties (see **Error! Reference source not found.**). The secret key is encoded in the quantum states of photons transmitted over the quantum channel. The authenticated classical channel is used to exchange measurement metadata between authorised users to remove errors and recover a shared key from the randomised 'raw key' through a process known as 'sifting'.

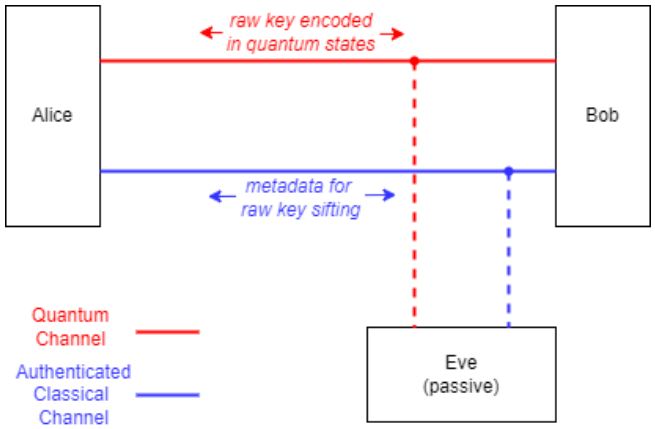


Figure 2: Block diagram showing a QKD system in a highly generalised form

QKD can be divided into the discrete-variable regime (DV-QKD) and the continuous-variable regime (CV-QKD); see (Hosseinidehaj et al., 2019) for a summary of the differences between the two regimes. In the discrete-variable regime, bits making up the raw key are encoded onto the discrete quantum states of a single photon, either by manipulating its polarisation or its phase. DV-QKD is typically realised using single photon sources and single photon detectors. By contrast, in the continuous-variable regime, the raw key is encoded in the quadratures of the optical field, which comprises many photons and can thus take on a continuous value in practical terms. CV-QKD is typically realised using either homodyne or heterodyne detectors at the receiver and promises a much higher potential key generation rate than DV-QKD, making it well-suited to handling the high volume of traffic through an IoT backhaul.

DV-QKD can be further divided into protocols that use a 'prepare-and-measure' approach and protocols that are based on quantum entanglement. In a 'prepare-and-measure' protocol – illustrated schematically in **Error! Reference source not found.** – Alice prepares a random sequence of quantum states. These are then transmitted over the quantum channel and Bob measures the quantum states to recover the raw key. In an entanglement-based protocol, a common source of entangled photon pairs distributes one photon each to Alice and Bob. The entanglement-based scheme obviates the need for random number generation in the coding and decoding process and allows Alice and Bob to use identical optical receiver architectures (Bedington et al., 2017).

Since the inception of the first QKD protocol in 1984, there has been a proliferation of protocols that aim to address the various weaknesses discovered in their predecessors, as the eternal cat-and-mouse game between attackers and defenders is played out in this new arena. Figure 2 depicts the evolution of some of the most significant QKD protocols since 1984 and categorises them according to whether they are continuous or discrete, and further whether they use a prepare-and-measure or entanglement principle. A summary of the major-milestone protocols follows.

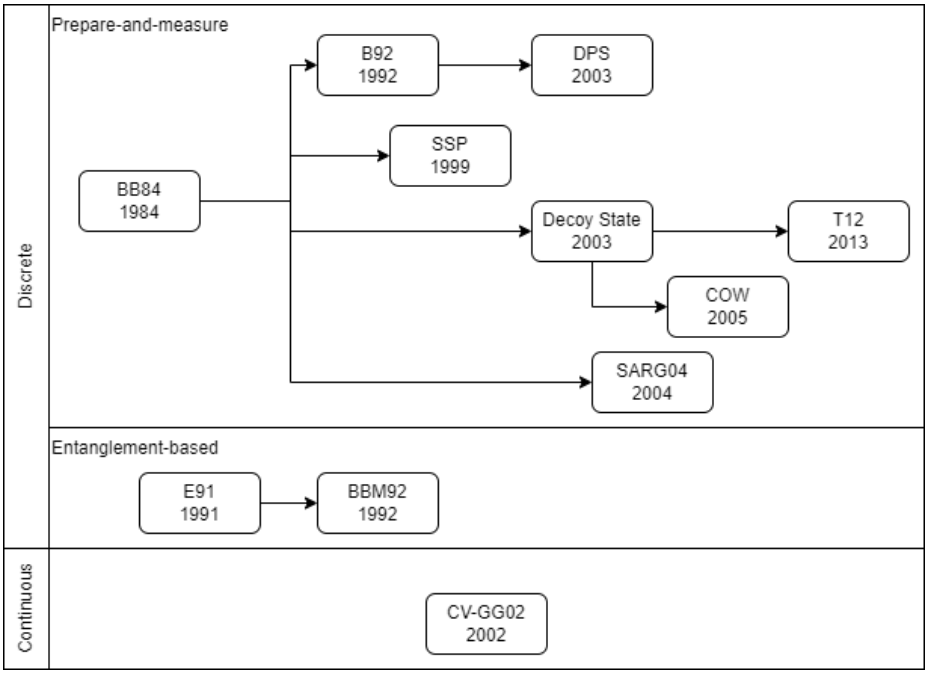


Figure 2: Classification and evolution of major-milestone QKD protocols.

The first QKD protocol to emerge was BB84, a prepare-and-measure type of DV-QKD, which encodes the raw key using the polarisation of single photons into one of four orthogonal states (Bennet and Brassard, 1984). E91 was the first protocol to use a common source of entangled photons, allowing Alice and Bob to use identical optical hardware (Ekert, 1991). The Decoy State protocol (Hwang, 2003) and SARG04 (Scarani et al., 2004) were developed in response to the vulnerability of BB84 to the photon-number splitting (PNS) attack. In a PNS attack, the attacker exploits the imperfections of real single photon sources. In general, single photon sources are imperfect, and sometimes emit more than one photon at a time. These clusters of photons share identical quantum states, and hence eavesdropper Eve can measure and recover information from one or more of the extraneous photons without violating the no-cloning theorem and without alerting Alice and Bob to her presence (Diamanti et al., 2016). This completes a non-exhaustive round-up of the watershed protocols in DV-QKD, but before moving on to the more esoteric subject of CV-QKD, it behoves us to provide a basic description of the mechanism by which DV-QKD achieves privacy against passive eavesdropping.

A single photon can exist in any one of a continuum of polarisation states, which can be thought of as the orientation of its electric field vector. When 'prepared'

- in a rectilinear basis, a photon becomes polarised either horizontally (encoding '0') or vertically (encoding '1');
- in a diagonal basis, a photon becomes polarised at either 45° (encoding '0') or 135° (encoding '1').

In the prepare-and-measure paradigm,

- Alice prepares a photon using a randomly selected polarisation basis, which is either rectilinear or diagonal, and transmits it to Bob.
- Bob then measures the photon using a randomly selected basis, which is either rectilinear or diagonal.

In the absence of Eve, if Bob happens to select the same basis as Alice's, then he can decode the photon's polarisation state into the right value (e.g., '0' for horizontal polarisation in the rectilinear basis); otherwise, Bob obtains the same value as Alice's at 0.5 probability. Now if Eve is interposed in the quantum channel and makes a measurement of the transmitted photon (see Figure 3), she may alter the photon's polarisation state such that Bob decodes it into a different value than Alice's even when he uses the same basis as Alice's.

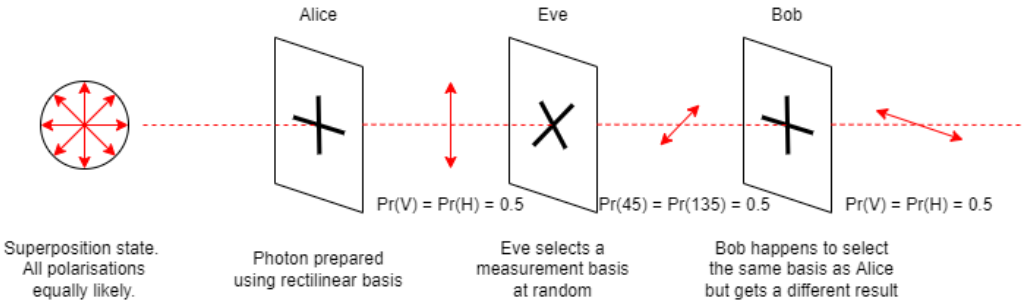


Figure 3: Passive eavesdropper detection in BB84.

Even when using the same basis, Alice and Bob obtain different values, which is an indication that the message may have been intercepted and hence altered by an eavesdropper.

Fundamentally, if a photon is first measured using the rectilinear basis, and then subsequently measured using the 'incompatible' diagonal basis, then the probability of finding the photon in any given state is 'reset'. Due to the Heisenberg uncertainty principle, the second, incompatible measurement disturbs the system in a way that creates maximum uncertainty of the measurement outcome (Wojcieszyn, 2022). This point is vital for understanding the security boon of QKD, because it means that an eavesdropper cannot intercept the raw key transmission without risking altering it in the process.

4. Application of QKD to Satellite IoT

Below, we assess each of the system parameters of QKD in turn in terms of potential performance and technical practicability.

Satellite access mode – direct vs indirect:
In a direct-access (a.k.a. direct-to-satellite) architecture, the satellite – typically in a LEO for the benefit of low latency – serves as the gateway, and the IoT devices communicate directly with the satellite. This provides the benefits of connecting geographically dispersed devices without the need for terrestrial infrastructure. This is particularly useful for ad hoc networks or geographically dispersed devices.

This requires the IoT device to have an onboard QKD module in addition to all the hardware necessary to establish and maintain an optical satellite connection, including (i) a telescope and (ii) a pointing, acquisition and tracking (PAT) system. The alternative, which is a far more plausible scenario for QKD, is an indirect-access scheme whereby IoT devices communicate with a terrestrial gateway, e.g., a cellular base station. The backhaul traffic from the base station is then carried via satellite.

See (Fraire et al., 2019) for a comparative analysis of the benefits and challenges associated with direct-to-satellite and indirect-to-satellite architectures for IoT networks.

QKD link direction – uplink vs downlink:
The raw key can be distributed via either an uplink or a downlink. A downlink is less lossy because beam wander occurs only over the last few km of the propagation path. Moreover, downlink QKD is less prone to denial-of-service attacks because the satellite only has a transmitter; an attacker aiming a high-intensity photon source at the satellite serves no purpose. It is also easier to build a large-aperture receiver (telescope) on a ground station than to deploy it in space. The principal advantage of the uplink scheme is that attackers have constrained physical access to tamper with an in-orbit receiver (Bedington et al., 2017).

Trusted Node relay:
Also known as a 'flying trusted node' architecture, the satellite establishes independent keys with Alice and Bob. Alice only knows the key she established with the satellite, and Bob only knows the key he established with the satellite. The general sequence is outlined as follows (Bedington et al., 2017):

1. Trusted relay establishes secure key with Alice, k_A.
2. Trusted relay establishes secure key with Alice, k_B.
3. Trusted relay combines the two keys: $k = k_A \oplus k_B$, where $\oplus$ denotes the exclusive-OR operation.

4. Parity announcement: the trusted relay distributes the combined key, k, to both Alice and Bob over the classical channel.
5. Alice recovers Bob's key from the combined key: $k_B = k_A \oplus k$.
6. Bob recovers Alice's key from the combined key: $k_A = k_B \oplus k$.
7. Alice can now use Bob's key to encrypt her plaintext for Bob, and vice versa.

Untrusted relays using quantum repeaters:
One of the fundamental problems with all forms of QKD, no matter the protocol used, is the rate-distance trade-off, which is well described in the literature. The key generation rate decays exponentially with distance, which indicates the need for a repeater chain to achieve global coverage (Takeoka et al., 2014). Whilst a conventional repeater in a communications network generally involves receiving, amplifying, and then re-transmitting a signal, the no-cloning theorem of quantum mechanics means that it is impossible for a repeater node to produce an exact copy of a quantum state for retransmission. One way around this problem is to establish a network of trusted node relays, which would decrypt the ciphertext and then establish a new key and re-transmit to the next node in the chain, but then these are no longer just repeaters, and it places additional security requirements on the space segment of the network. The ideal solution would be to use a network of untrusted repeaters, and recent developments in quantum physics research suggest that this may be feasible and even practicable in the near future (Mastriani and Iyengar, 2020). Quantum repeaters are based on the principle of *entanglement swapping* (see Definition 1), which currently is dependent on the development of quantum memory to overcome the problem of asynchronous photon arrival time, but this is a highly active area of research and progress is being made (Takeoka et al., 2014).

Definition 1: *Entanglement swapping* is a process whereby a pair of independently generated photons can be entangled so that they have correlated quantum states (Pan et al., 1998). This process can be repeated multiple times to effectively extend the range of a quantum transmission system, and is the key principle behind quantum repeater technology and for overcoming the rate-distance trade-off.

Untrusted relays using MDI-QKD:
In an alternative approach to realising an untrusted relay network architecture, namely a measurement-device-independent QKD (MDI-QKD) protocol, a third party (Charlie), who need not be trustworthy, is responsible for making all measurements on the quantum state of the encoded raw key (Hosseinidehaj and Malaney, 2017). Alice and Bob both prepare randomised quantum states using pre-agreed bases, and transmit them to Charlie, who measures the correlations between the signals from Alice and Bob, and then announces the correlations over an unsecure channel. If the correlation is high, Alice and Bob know that they have agreed on the same bit, but nobody else knows what that bit is; not even Charlie, who can only measure the superposition of Alice and Bob's signals but is unable to separate them.

Continuous-variable vs discrete-variable protocols:

All DV-QKD schemes suffer from the problem of losing photons over free-space transmission through the atmosphere. Dust and cloud cause loss of photons (Mastriani and Iyengar, 2020), severely limiting the key generation rate. DV-QKD protocols are better suited to terrestrial fibre transmission than long-distance free-space communications because loss of photons in the former is appreciably less. Moreover, when single-photon detectors are used, there is a certain amount of 'dead time', which can significantly hold back the maximum achievable baud rate (Djordjevic, 2020a). Thus, we propose that the satellite-integrated IoT infrastructure of the future will require the adoption of CV-QKD protocols to cope with backhaul demands.

Key Generation – synchronous vs stockpiled:
In the synchronous scheme, the connection is established on demand, and the secret key is generated in real time. This most closely resembles the format of present-day public-key infrastructure, where keys are generated on demand. However, this would work only if the secret key generation rates were sufficient to handle real-time IoT bandwidth requirements, and present demonstrations fall far short of this. Synchronicity also requires implementation of a globe-spanning constellation with either a relay of trusted nodes or untrusted repeaters.

The alternative is an asynchronous scheme, whereby a one-time pad is (i) accumulated whenever line-of-sight over a ground station is available and (ii) stockpiled for later user. This could in principle be realised using only a

single CubeSat in polar orbit to distribute a secure key using QKD to any two points on the globe (Sidhu et al., 2022). Keys are retrieved from storage and used for encryption/decryption as needed.

In Table 1, we synthesise two possible solutions for a globe-spanning QKD-secured satellite-integrated IoT network. The first is an ultimate solution (red dots ●) for implementing QKD security for satellite-integrated IoT backhaul – although this one might be many years off due to the nascent nature of certain technologies. And in the other (blue dots ●) we propose a solution that might be more attainable in the near-term and begin to satisfy the potentially urgent need for quantum-resistant cryptography.

Table 1: QKD architecture synthesis summary.

Parameter	Option 1	Option 2
Satellite access method	Direct-to-satellite	Indirect ● ●
Space segment security	Trusted relay ●	Untrusted ●
QKD link direction	Uplink	Downlink ● ●
Key generation	Synchronous ●	Stockpiled ●
Encoding basis	Discrete variable (DV) ●	Continuous variable (CV) ●

Several experiments have been carried out in recent years that lend credence to the concept of a satellite-based QKD network supporting an indirect-access architecture. All three of the examples listed below use the parameter set indicated by the blue dots (●) in Table 1.

In the first full-scale practical demonstration of satellite-based QKD, Liao et al. (2017) showed that a sat-to-ground data link can be established with a 1 kbps data rate over a range of 1200 km, during the testing of their satellite Micius in 2015. The window of effective key transfer was 273 seconds, and at a single fly-over per night, this resulted in a total key rate of 273 kb/night/satellite. Although this experiment was limited to only one node in the space segment, one can imagine extending this technology to a globe-spanning satellite constellation, where the inter-satellite links are secured using QKD.

Modelling performed by Mazzarella et al. (2020) as part of the QUARC project showed that a constellation of 15 CubeSats would be sufficient to provide a QKD-secured service to the UK with reasonable availability and key generation rate. They modelled a trusted relay network using Decoy State DV-QKD and managed to achieve a best-case key generation rate of 300 kb/night/satellite. This was based on a highly conservative assumption that interference prohibits transmission during the daytime or twilight hours.

Tiangong-2 Space Lab is a Chinese space station orbiting in a LEO at medium inclination. It has been used as a testbed for a trusted relay for a QKD experiment, achieving a sifted key generation rate of 1.33 Mb per day under best-case conditions (Li et al., 2022). This experiment showed that by combining satellites in sun synchronous orbit (SSO) and medium inclination orbit (MIO), a higher overall key rate per day can be achieved due to the multi-pass benefit of MIOs. A sun-synchronous orbit is a polar orbit (approximately 90° inclination) with a precession synchronised to the revolution of the earth around the sun, so that a satellite in SSO can be both exposed to the sun at all times for power generation, and have line-of-sight to the dark side of the Earth for clarity of optical communications at the same time.

Table 2 summarises the important performance figures achieved in the experiments discussed above.

Table 2: Summary of recent milestones in satellite-based QKD performance.

Experiment	Year	Location	Orbital Height (km) at zenith	Best Key Rate (kb/sat/night)	Reference
Micius (QUESS)	2017	Xinglong, China	500	300	Liao et al., 2017
QUARC	2020	Southeast England	574	300*	Mazzarella et al., 2020
Tiangong-2	2022	Nanshan, China	400	1330	Li et al., 2022

*The QUARC result is based on modelling and simulation rather than real-world experimental data.

5. Concluding remarks

We have established the need and urgency for PQC and provided a light introduction to QKD in its multitudinous forms. The security of current cryptosystems such as RSA is predicated on the absence of efficient solutions to some computational problems, but efficient solutions may become available with or without quantum computing. In contrast, QKD leverages quantum-mechanical properties to achieve information-theoretic

security. Having explained some of the variations in architectural design that are possible within QKD, we have provided some recent examples of practical demonstrations of the technology that lend credence to its viability for securing IoT backhaul over satellites in the future. We proposed two possible architectures for satellite-integrated IoT using QKD, one near-term and one longer-term. The ideal QKD architecture from a security perspective is one involving a relay of untrusted quantum repeaters establishing point-to-point secure keys between terrestrial IoT gateways over an optical (quantum) channel. Synergy among pre-quantum cryptography, PQC and QKD are required to achieve security against classical and quantum adversaries. Side-channel attacks are applicable to all cryptographic paradigms and hence comprehensive vulnerability assessments must accompany advances in QKD implementation (Sajeed et al., 2021).

References

Alagic, G., Apon, D., Cooper, D., et al. (2022) 'Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process', NISTIR 8413-upd1. DOI: 10.6028/NIST.IR.8413-upd1.

Azuma, K., Mizutani, A. and Lo, H.-K. (2016) 'Fundamental rate-loss trade-off for the quantum internet', *Nature Communications*, vol. 7, no. 1, p. 13523. DOI: 10.1038/ncomms13523.

Bedington, R., Arrazola, J. M. and Ling, A. (2017) 'Progress in satellite quantum key distribution', *npj Quantum Information*, vol. 3, no. 1, p. 30. DOI: 10.1038/s41534-017-0031-5.

Bennett, C. H. and Brassard, G. (1984) 'Quantum cryptography: Public key distribution and coin tossing', *Theoretical Computer Science*, vol. 560, pp. 7–11. DOI: 10.1016/j.tcs.2014.05.025.

Bernstein, D. J., Hülsing, A., Lange, T. and Rekleitis, E. (2022) 'Post-quantum cryptography: integration study', European Union Agency for Cybersecurity. Available at https://www.enisa.europa.eu/publications/post-quantum-cryptography-integration-study/@@download/fullReport.

Bernstein, D. J. and Lange, T. (2017) 'Post-quantum cryptography', *Nature*, vol. 549, no. 7671, pp. 188–194. DOI: 10.1038/nature23461.

Cao, Y., Zhao, Y., Wang, Q., et al. (2022) 'The Evolution of Quantum Key Distribution Networks: On the Road to the Qinternet', *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 839–894. DOI: 10.1109/COMST.2022.3144219.

Castryck, W. and Decru, T. (2022) 'An efficient key recovery attack on SIDH (preliminary version)', Cryptology ePrint Archive, Paper 2022/975. Available at https://ia.cr/2022/975.

Centenaro, M., Costa, C. E., Granelli, F., et al. (2021) 'A Survey on Technologies, Standards and Open Challenges in Satellite IoT', *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1693–1720. DOI: 10.1109/COMST.2021.3078433.

Cheng, C., Lu, R., Petzoldt, A. and Takagi, T. (2017) 'Securing the Internet of Things in a Quantum World', *IEEE Communications Magazine*, vol. 55, no. 2, pp. 116–120. DOI: 10.1109/MCOM.2017.1600522CM.

Commonwealth of Australia. *Archives Act 1983:* C2021C00366 (No. 79, 1983) Canberra, Office of Parliamentary Counsel.

Diamanti, E., Lo, H.-K., Qi, B. and Yuan, Z. (2016) 'Practical challenges in quantum key distribution', *npj Quantum Information*, vol. 2, no. 1, p. 16025. DOI: 10.1038/npjqi.2016.25.

Djordjevic, I. B. (2020a) 'Hybrid QKD Protocol Outperforming Both DV- and CV-QKD Protocols', *IEEE Photonics Journal*, vol. 12, no. 1, pp. 1–8. DOI: 10.1109/JPHOT.2019.2946910.

Djordjevic, I. B. (2020b) 'Secure, Global Quantum Communications Networks', *22nd International Conference on Transparent Optical Networks (ICTON)*, IEEE, pp. 1–5. DOI: 10.1109/ICTON51198.2020.9203116.

Ekert, A. K. (1991) 'Quantum cryptography based on Bell's theorem', *Phys. Rev. Lett.*, American Physical Society, vol. 67, no. 6, pp. 661–663. DOI: 10.1103/PhysRevLett.67.661.

ETSI (2018) 'Quantum Key Distribution (QKD); Vocabulary', Group Report reference DGR/QKD-0007_Ontology V1.1.1. https://www.etsi.org/deliver/etsi_gr/QKD/001_099/007/01.01.01_60/gr_qkd007v010101p.pdf.

Fernández-Caramés, T. M. (2020) 'From Pre-Quantum to Post-Quantum IoT Security: A Survey on Quantum-Resistant Cryptosystems for the Internet of Things', *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6457–6480. DOI: 10.1109/JIOT.2019.2958788.

Fraire, J. A., Céspedes, S. and Accettura, N. (2019) 'Direct-To-Satellite IoT - A Survey of the State of the Art and Future Research Perspectives: Backhauling the IoT Through LEO Satellites', in *Ad-Hoc, Mobile, and Wireless Networks*, Lecture Notes in Computer Science, Springer International Publishing, vol. 11803, pp. 241–258. DOI: 10.1007/978-3-030-31831-4_17.

Gidney, C. and Ekerå, M. (2021) 'How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits', *Quantum*, Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften, vol. 5, p. 433. DOI: 10.22331/q-2021-04-15-433.

Grover, L. K. (1996) 'A fast quantum mechanical algorithm for database search', *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing - STOC '96*, ACM Press, pp. 212–219. DOI: 10.1145/237814.237866.

Hosseinidehaj, N. and Malaney, R. (2017) 'CV-MDI Quantum Key Distribution via Satellite', *Quantum Info. Comput.*, vol. 17, no. 5–6, pp. 361–379. DOI: 10.26421/QIC17.5-6-1. Available at 10.48550/ARXIV.1605.05445.

Hosseinidehaj, N., Babar, Z., Malaney, R., et al. (2019) 'Satellite-Based Continuous-Variable Quantum Communications: State-of-the-Art and a Predictive Outlook', *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 881–919. DOI: 10.1109/COMST.2018.2864557.

Hwang, W.-Y. (2003) 'Quantum Key Distribution with High Loss: Toward Global Secure Communication', *Physical Review Letters*, vol. 91, no. 5, p. 057901. DOI: 10.1103/PhysRevLett.91.057901.

Liao, S.-K., Yong, H.-L., Liu, C., et al. (2017) 'Long-distance free-space quantum key distribution in daylight towards inter-satellite communication', *Nature Photonics*, vol. 11, no. 8, pp. 509–513. DOI: 10.1038/nphoton.2017.116.

Li, Y., Liao, S.-K., Cao, Y., et al. (2022) 'Space–ground QKD network based on a compact payload and medium-inclination orbit', *Optica*, vol. 9, no. 8, p. 933. DOI: 10.1364/OPTICA.458330.

Mastriani, M. and Iyengar, S. S. (2020) 'Satellite quantum repeaters for a quantum Internet', *Quantum Engineering*, vol. 2, no. 4. DOI: 10.1002/que2.55.

Mazzarella, L., Lowe, C., Lowndes, D., et al. (2020) 'QUARC: Quantum research Cubesat—A constellation for quantum communication', Cryptography, vol. 4, no. 1. DOI: 10.3390/cryptography4010007.

Nordholt, J. E., Hughes, R. J., Morgan, G. L., et al. (2002) 'Present and future free-space quantum key distribution', Mecherle, G. S. (ed), San Jose, CA, pp. 116–126. DOI: 10.1117/12.464085.

Pan, J.-W., Bouwmeester, D., Weinfurter, H. and Zeilinger, A. (1998) 'Experimental Entanglement Swapping: Entangling Photons That Never Interacted', *Physical Review Letters*, vol. 80, no. 18, pp. 3891–3894. DOI: 10.1103/PhysRevLett.80.3891.

Sajeed, S., Chaiwongkhot, P., Huang, A., et al. (2021) 'An Approach for Security Evaluation and Certification of a Complete Quantum Communication System', *Scientific Reports*, vol. 11, no. 1, p. 5110. DOI: 10.1038/s41598-021-84139-3.

Scarani, V., Acín, A., Ribordy, G. and Gisin, N. (2004) 'Quantum Cryptography Protocols Robust against Photon Number Splitting Attacks for Weak Laser Pulse Implementations', *Physical Review Letters*, vol. 92, no. 5, p. 057901. DOI: 10.1103/PhysRevLett.92.057901.

Shor, P. W. (1994) 'Algorithms for quantum computation: discrete logarithms and factoring', *Proceedings 35th Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, pp. 124–134. DOI: 10.1109/SFCS.1994.365700.

Sidhu, J. S., Brougham, T., McArthur, D., et al. (2022) 'Finite key effects in satellite quantum key distribution', *npj Quantum Information*, vol. 8, no. 1, p. 18. DOI: 10.1038/s41534-022-00525-3.

Takeoka, M., Guha, S. and Wilde, M. M. (2014) 'Fundamental rate-loss tradeoff for optical quantum key distribution', *Nature Communications*, vol. 5, no. 1, p. 5235. DOI: 10.1038/ncomms6235.

Wojcieszyn, F. (2022) *Introduction to Quantum Computing with Q# and QDK*, Springer Nature. DOI: 10.1007/978-3-030-99379-5.

Wootters, W. K. and Zurek, W. H. (1982) 'A single quantum cannot be cloned', *Nature*, vol. 299, no. 5886, pp. 802–803. DOI: 10.1038/299802a0.

Zhou, I., Makhdoom, I., Shariati, N., et al. (2021) 'Internet of Things 2.0: Concepts, Applications, and Future Directions', *IEEE Access*, vol. 9, pp. 70961–71012. DOI: 10.1109/ACCESS.2021.3078549.

Zou, Y., Zhu, J., Wang, X. and Hanzo, L. (2016) 'A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends', *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1727–1765. DOI: 10.1109/JPROC.2016.2558521.

Commentary on Healthcare and Disruptive Innovation

Hilary Finch[1,2], Affia Abasi-Amefon[3], Jung, Woosub[4], Lucas Potter[5] and Xavier-Lewis Palmer[1,2,5]

[1]Old Dominion University, Norfolk, USA
[2] CySecSol, Franklin, USA
[3] University of Tartu, Tartu, Estonia
[4]William & Mary, Williamsburg, USA
[5] BiosView, Oswego, USA

hfinc003@odu.edu
amefon.affia@ut.ee
wjung01@wm.edu
lpott005@odu.edu
xpalm001@odu.edu

Abstract: Exploits of technology have been an issue in healthcare for many years. Many hospital systems have a problem with "disruptive innovation" when introducing new technology. Disruptive innovation is "an innovation that creates a new market by applying a different set of values, which ultimately overtakes an existing market" (Sensmeier, 2012). Modern healthcare systems are historically slow to accept new technological advancements . This may be because patient-based, provider-based, or industry-wide decisions are tough to implement, giving way to dire consequences. One potential consequence is that healthcare providers may not be able to provide the best possible care to patients. For example, if a healthcare provider does not adopt new technologies or approaches to medical treatment, they may not be able to offer the same level of care as a provider who has embraced those innovations. This leads to lower quality of care and poorer patient outcomes. Another consequence is that healthcare providers who do not adapt to disruptive innovations may lose market share to competitors who are more forward-thinking and willing to embrace new technologies and approaches. This can harm the provider's financial performance and sustainability.Not adapting to disruptive innovations in healthcare can result in missed opportunities to improve the efficiency and effectiveness of medical treatment. If a healthcare provider does not adopt electronic medical records, they may miss out on the benefits of faster and more accurate information sharing, improving patient care. Once the decision to implement technology in a specific healthcare industry is made, concerns about patient safety, an aversion to change, and hospital-wide compliance with regulations begin to arise (WynHouse, nd.). The healthcare technology industry also boomed with the COVID-19 outbreak. The COVID-19 outbreak has led to significant advancements and innovations in medical technology. In order to diagnose, treat, and prevent the spread of the virus, healthcare providers and researchers have had to develop and deploy new technologies and approaches. The COVID-19 outbreak has highlighted the importance of the medical industry and the essential role it plays in society. This has led to increased funding, support for medical research and development, as well as a greater appreciation for the work of healthcare providers. This has created opportunities for growth and innovation in the medical industry. It also placed enormous strain on global health systems, disrupting healthcare by increasing the risk of fraud and deception; the risk of hospital operations and assets being compromised, disrupted, or altered; and the increased use of telehealth resulting in a breakdown between providers and consumers (Kuehn,2021). This article will cover the effects/impact of disruptive innovations/technologies introduced into healthcare industries over the short term through a light review of disruptions and responses, followed by commentary and policy recommendations.

Keywords: Healthcare, Disruption, Cybersecurity, Cyberbiosecurity, Biocybersecurity, Disruptive Innovation,

1. Background

For many years, the misuse of technology has plagued the healthcare system. "Disruptive innovation" offers an issue for healthcare systems, particularly in introducing new technologies. Historically, healthcare systems have developed or gained essential technology advances slower than other industries. The sluggish pace may be due to the difficulty of implementing patient-based, provider-based, or industry-wide technological decisions (Thimbleby,2013). Medical technology has the potential to significantly improve patient outcomes, but also carries risks. By taking a more cautious approach to introducing new technologies, healthcare providers can ensure that any new innovations are thoroughly tested and proven to be safe before they are widely adopted. Not all patients have equal access to healthcare, and the rapid introduction of new technologies can exacerbate existing disparities. By taking a more deliberate approach to introducing new technologies, healthcare providers can ensure that all patients have access to the care they need. The healthcare system is a complex and interconnected system, and the introduction of new technologies can have unintended consequences. By slowing the pace of technological change, healthcare providers can consider the long-term impacts of new innovations and ensure that they are sustainable over the long term. Further, if stakeholders decide to deploy technology in particular healthcare businesses, questions regarding patient safety, resistance

to change, and hospital-wide regulatory compliance arise (WynHouse, nd.). Although the medical instruments, drugs, procedures, and systems developed to improve people's health and quality of life are now simpler to use due to recent advancements in medical technology, their application across multiple departments remains challenged. The impact of this problem blunts the possibility of hospitals exploiting the near full potential of their technology, let alone their security technology. This brings into question if there is a disruptive innovation to be had at all.

The phrase "disruptive innovations" (DI) is new to the healthcare field. Because we now live in a digital age, healthcare should adapt its approach to technology, which should involve modifying how medical treatment is provided. If we do not adapt to healthcare technology, it could result in a lack of equitable access to medical care. This is because technology can play a crucial role in improving the accessibility of healthcare, particularly in underserved and remote communities. For example, the use of telemedicine and remote monitoring systems can allow healthcare providers to reach patients who may not otherwise have access to medical care.

Failing to adapt to healthcare technology can also result in preventable illnesses spreading to larger populations. This is because technology can help identify and track outbreaks of infectious diseases, and facilitate the rapid sharing of information and resources to contain and control the spread of these diseases. Not using technology to its full potential, we may miss out on opportunities to identify and address potential public health threats before they become widespread. Overall, the use of technology in healthcare can help to improve access to care and prevent the spread of preventable illnesses.

The disruption caused by DI leads to tremendous change, causing a ripple effect on how things are done across the business. Healthcare-specific disruptive innovations are not limited to devices or digital technologies but also cover innovations in basic science, diagnostics (i.e., pathological or radiological diagnostic discoveries), education, processes (i.e., novel health policies, new processes or roles like the introduction of nurse practitioners or patient assistants), and in technique (i.e., the introduction of novel medical techniques within a specialty) (Sounderajah et al.,2021). However, we focus our analysis on devices and digital technologies to explore the security ramifications of introducing these technologies into the healthcare industry.

1.1 Healthcare DI, types of healthcare DI, the scope of work

Popular examples of technological disruptions in the healthcare industry cover software applications, (i.e., mHealth applications, telehealth, patient data systems, and clinical decision support systems), devices, wearables, monitors, imaging, system infrastructure innovations through cloud computing, IoT, and blockchain, and other applied innovations using RFID , AR/VR, and AI The software applications such as the CDSS examine data contained within electronic health records to provide clinicians and other healthcare professionals with reminders and prompts to assist patients in receiving evidence-based guidelines, ultimately leading to improved care and records. (CDC, 2022). With applied innovations, AV/VR has been used to promote patient rehabilitation (Brooks,2014), while RFID makes it simpler for providers to track, communicate, and identify resources and patients (Haddara & Staaby, 2018). Recent developments in the healthcare industry have also brought on the necessity to implement AI in the healthcare context. AI can reduce healthcare costs by reducing staff burnout, reducing patient wait times, and tackling disease complexity(Bellucci,2002). These entail innovations in the form of using AI assistants to grade and sort patient images and text data, being able to address or at least route patient concerns through natural language processing modalities, to examine literature, patient data, and health care provider inputs to either decipher patient pathologies or solution(s) to patient pathologies and reduce harms and lawsuits that would result, among other innovations (Wynants et al., 2020; Shaheen,2021). Reflecting on benefits of these disruptive innovations and the move towards adoption, it is essential to consider the security implications of these proposed innovations, post-adoption, especially with the impact of COVID-19 on the industry.

According to Sounderajah et al. (2021), the concept of "disruptive innovation" has made its way into health care. The article discusses the outcome of a systematic review, characterizing and categorizing the concept of “disruptive innovation” into basic science, device, diagnostics, digital health, education, processes, and technique. The paper also discussed great interest in technology-related innovations such as "omics” technologies, mobile health applications, telemedicine, and health informatics. However, the authors highlight that the use of disruptive innovation remains inconsistent, and its current definition does not accommodate exploration, especially for policymakers.

1.2 Related Work with process considerations in Healthcare DI

The main aim of this study is a commentary and light examination of the notion of "disruptive innovation" in healthcare within the scope of cybersecurity. Sensmeier claims that healthcare employees see the growth of technological innovation in their daily lives (Sensmier,2012). The use of smart patient care equipment, electronic patient care files, blood glucose meters, and portable gadgets in healthcare disruption can be a vital tool for constructing a higher quality, more convenient, and lower cost healthcare system. In this study, Sensmeier claims that advances in healthcare technology will allow people to experience faster, more individualized care (Sensmier, 2012).While technology is improving and new ideas are being introduced, the healthcare system is becoming more complicated, inefficient, and stressful for practitioners (Harris,2018). This article supports the idea that every organization's culture should encourage strong technical communication and coordination among clinicians worldwide. Still, the infrastructure needs to be able to support new ideas that improve workflow and make it easy to use relevant technical or institutional knowledge at the point of care. By leveraging disruptive innovation, incorporating evidence into care delivery and decision-making, involving patients and families in healthcare decisions, and improving care coordination across organizations, we will accelerate progress toward improving our nation's health.

Yellowlees et al. (2011) state that asynchronous medicine would thrive within a facilitated network model, which would change the jobs of most of the people involved in caring for a patient. The facilitated network model would consist of an electronic system through which healthcare professionals and patients share information, like transmitting healthcare records. This practice is formally known as "precision medicine." Precision medicine includes the delivery of patient-specific, focused treatment. This is accomplished by utilizing electronic health records (EHRs) to enhance patient outcomes.Yellowlees et al. (2011) asserts that while asynchronous medicine may be well-established and advantageous in specialties like radiology, there is little research regarding the benefits of asynchronous medicine in areas that traditionally rely on physical doctor-patient interaction. A transition to asynchronous medicine would significantly change the roles of primary care doctors, specialists, patients, and nonmedical providers. It states that putting asynchronous data sets into patient records would make it easier to compare patients over time and between groups and make it possible to get second opinions without requiring excessive and unnecessary travel from the patient. While this may make healthcare optimal, we could open ourselves up to future vulnerabilities. It is reportedly common for radiologists to work remotely (Yellowees et al, 2011). Patient data is sent to these outside of network workstations. However, after reviewing previous work evaluating disruptive innovation in healthcare, we find that cybersecurity aspects of these innovations in their development and adoption by the healthcare industry remain under-examined. Due to the benefits of disruptive innovations, we must raise cybersecurity concerns that harm healthcare services and patients.

2. Healthcare DI pre-pandemic, security considerations

Before the pandemic, there were already many specialized hospital information systems, like electronic health record (EHR) systems, e-prescribing systems, practice management support systems, clinical decision support systems, radiology information systems, and computerized physician order entry systems. These information systems, widely used by many hospitals and other healthcare facilities, are highly specialized and collect, manipulate, process, store, and depend on patients' personal and sensitive information. There are several different types of EHR systems that are utilized in the healthcare setting, and they have their own particulars in security needs (Adracare, nd.). As the healthcare sector has more potential for harm than nearly any other industry, cybersecurity should be a priority alongside DI, but it is not (Schulman et al,2009). Alarmingly, Jalali and Kaiser (2018) noted that "70% of hospital boards include cybersecurity in their risk management oversight, and only 37% of hospitals perform annual incident response exercises." Just in 2021, there have been 108 individualized ransomware attacks that affected 2,302 medical facilities. Within these facilities, it was estimated that 19.76 million patient records were impacted. Thes attacks cost the medical facilities $7.8 billion in downtime (Bischoff,2021). These indicate major gaps in hospital security oversight and help explain numerous hospital attacks. Hospital organizational culture had insufficient threat and vulnerability awareness or preparation. Further, they cannot guarantee the availability of resources for the product they promote or use. This problem has been long-running. Garret et al. (2006) found in the analysis of rural hospitals that the drive to use technology to address US healthcare problems was hampered through sub-optimal utilization; at the time, less than 10% of hospitals had an electronic medical record with an estimated less than 20% of healthcare providers utilizing complete computerized patient records. This has improved, according to Healthit.gov, which stated, "In data from 2019 and 2021, 86% of non-Federal general acute care hospitals had adopted a 2015 Edition certified

electronic health record (EHR). In contrast, only 40% of rehabilitation hospitals and 23% of specialty hospitals had adopted a 2015 Edition certified EHR," (*Adoption of electronic health records by hospital service type,2019-2021*). Nonetheless, hospitals have far to go in optimally using electronic records.

3. Barriers associated with data-sharing with DI technologies

DI technology definitionally requires adjustment. Lack of policy, norms, or quality may be technical. To meet the healthcare industry's changing needs, software, installation, upgrades, and equipment optimization cost money. (AAP, nd). The presence of a competitive market, networks, and systems that make it easy to exchange data and personally identifiable information at risk can all act as trust obstacles. This might be because of the reliance on vendors and market providers, who have relaxed regulations regarding data protection, causing a person authorized to access protected health information at a covered entity to disclose the information inadvertently due to a breach of security at the covered entity (HIMSS, nd).

4. Malware of Various Types, Including Ransomware

There are a variety of additional forms of malware, like ransomware, that pose a threat to healthcare providers and organizations. Examples include ransomware, Trojans, and phishing. Phishing is particularly effective since it is directed at the specific user and attempts to dupe the user into giving sensitive information, clicking on a malicious link, or opening an infected file using social engineering techniques. Emails sent using spear phishing to target specific individuals have a higher success rate than emails sent using broader tactics. When it comes to cybersecurity in the healthcare industry, a comprehensive incident response strategy is required to ensure that potential security breaches are prevented or dealt with promptly and effectively. This is to protect the patient's personal information. For instance, an assault was launched against the information technology department at the University of Vermont Medical Center as recently as October 2020. During the nearly one month without EHR (electronic health records) and other essential medical technology, the staff at UVMC worked using paper tools. The resolution of this assault took over a month, and even though the company chose not to pay the ransom but instead to shut down its networks, the estimated cost of the attack was still around fifty million dollars (AAMC, nd).

These factors might cause providers to struggle, including the cost to the business, the time spent learning new software, the legality of the software and its data collection and restrictions, the fear of learning new ways of doing things, the usefulness of the product, and the complexity of implementing the new DI technology. Interactions between the device and the user are referred to as "human factors" in medical device equipment. Everything from the device's size, shape, and controls to how intuitive it is to use and how well the instructions are written contributes to the device's usability. Human factors in medical device design focuses on making sure the device isn't dangerous to the user and can be operated with minimal effort and time loss.

DI software and technology are not designed with the provider in mind. Some of this results in real or perceived increased effort for personnel. Rural medical facilities should be taken into account because they have unique problems with adopting DI because they don't have enough staff, space, or reliable internet access. According to a 2018 FCC estimate, 31% of rural households in the United States still do not have access to broadband internet (Eighth Broadband Progress Report, nd). Some DI decisions also disrupt patient life owing to the complexity of the program and its applications, a lack of access to information or insufficient information available to the patient, or the technology itself being limited in its use. Cultural issues for digital services include telehealth, telemonitoring, technological skepticism, digital literacy, the expense of required hardware to operate the program, and device compatibility. To take a page from Garrett et al. (2006), patients must understand the privacy and security policies involved in their care, along with how to keep their information private and secure.

5. Healthcare DI during and post-pandemic, security considerations

Covid-19's healthcare technology expansion boom placed enormous pressure on health systems globally, disrupting healthcare services by raising the risk of fraud and deception, altering and interrupting cyberbiosecurity, and creating a breakdown between providers and consumers through the increased use of telehealth (Kuehn, 2021). The pandemic gave perfect conditions for a spike in ransomware attacks on healthcare institutions like the Brno University Hospital in the Czech Republic, the World Health Organization, and other healthcare supply chains (He et al., 2021). Ransomware attacks within an industry would increase in frequency and pay, making the healthcare system even more fragile than before the pandemic.And as healthcare companies improve their cybersecurity defenses, criminals adopt new modus operandi to maximize financial

gain by executing more complex attacks within a single breach, often leading to a complete system shutdown (Lang et al.,2021). Tracking systems, technology-based social distancing, mobile app-based contact tracing systems, telemedicine consultations, e-quarantine technology, AI-based patient care, QR code check-in systems, and AI-based temperature scanning are all examples of innovations in the healthcare sector that resulted from the impact of COVID-19 (Liu et al.,2022). Various interviews with hospital professionals and managers revealed that, subsequently of the COVID-19 epidemic, digital care has increased (Lee et al.,2021). The pandemic encouraged the need for hybrid healthcare services to deliver continuous enhanced service thanks to the innovation of digital technology (Liu et al. 2022). The COVID-19 pandemic was also a large factor in facilitating the growth of digital health technology (DHT). DHT is the fusion of digital technologies with society, health, and medical care. With the prioritization of DHT to “flatten the curve," DHT innovations are at the forefront of DI technology (Whitelaw et al., 2021). These innovations are changing how we treat and deliver patient care. As discussed in the previous sections, medical technology was already slow to implement, and the pandemic brought the same doubts and fears, widening the gap between technology and providers.

6. In the scope of COVID-19 and cyberbiosecurity

Many people are concerned with the amount of information and data that must be sacrificed to receive quality care. While these were concerns before the pandemic, the onset of COVID-19 increased the necessity of securing digital resources due to the increased threat of cyber-attacks by threat actors (Adler et al.,2021).

Contact-tracing Privacy is about tracking and identifying people who have had contact with a person or people who are sick. This method of identifying and monitoring individuals who may or may not be sick has been around for centuries (Contact Tracing, nd). In past years, contact tracing was a manual process that was limited in identification and notification within a quality time frame to cause the least harm. Notably, the past infectious diseases that were traced were for infections of less than 10,000 people worldwide, whereas COVID-19 has infected 619 million as of this writing (WHO,2022), proving the need for other options for contact tracing. The use of cell phones for contact tracing began innocently enough; it was a quick, efficient, and discretionary way to spread COVID-19 infection information. In the age of technology, most people’s personal information is stored on their devices: banking services, biometrics, credentials, and access to healthcare portals with known vulnerabilities. By allowing countries to use or access location services of smart devices in an attempt to contact trace, the end user is opening themselves to security threats and vulnerabilities leading to unauthorized access to personally identifiable information. An alternative method of contact tracing with more privacy is the use of Bluetooth technology in place of location services. Bluetooth collects significantly less information; some of it is voluntary, like phone numbers. This technology detects if you have been close to an infected person(s) and notifies you only if you have been in the same area as the infected individual without invading the infected individual's privacy, which is useful for the general public. It hinders the contact tracing process, as the entire connection and divulging of PII is completely voluntary, leaving it open to human error and false reporting. The COVID-19 pandemic showed how important it is to keep track of disease data for researchers, policymakers, and the general public. The disease data includes the number of cases, the results of diagnostic tests, and information about general trends, which makes it a target for malicious actors.Diagnostic testing and counting are highly valuable and used to measure epidemic levels, making it potentially catastrophic if the surveillance data is compromised (Adler et al.,2021).

7. Policy recommendations and solutions to security challenges with Healthcare DI (pre during-post pandemic)

Taking focused facilities and user networks out of the current mixed healthcare delivery models could save money and be more efficient. Still, it could also break up the way care is given. Care coordination can also be done differently by a patient-centered medical home, telephonic services, web-based decision-making software, and personal health records. But it is important to remember that the healthcare system comprises many business models that depend on each other. Policymakers in health care need to be aware of the hidden cost of supporting and renewing rules that slow down innovation in the long run. With less money, hospitals and doctors have to work even harder to meet their value proposition of providing complex and expensive medical care. They are even less likely to give work to businesses that add value (Hwang et al.,2008). The topic of cyberbiosecurity is very multidisciplinary, and it will be beneficial to integrate the skills and methodologies already in use from various sectors with the requirements posed by the life sciences. Several other cyber applications have been proposed as potential solutions to some emerging problems regarding cyberbiosecurity. Putting them into reality might be difficult because even the most fundamental concepts of information security

require improved adaptation to the framework of bioscience. Similarly, it will be necessary to hone and expand the traditional CIA triad (confidentiality, integrity, and availability) and extend the ideas made earlier to match them properly with the current requirements. Traditional cybersecurity concerns are not always relevant to all of the newly discovered difficulties. Therefore, it will be essential to figure out which difficulties the currently available cyber-approaches could or could not discover or guard against. There is a need for a rethinking of quality assurance testing since the completed product or system may not appear exactly like what was stated (in digital form). Quality control and assurance measurements only show how a biological system was and is. They cannot predict the future or provide enough information for cyber defense. Safeguarding policies should consider the unique qualities of "information" in the biological sciences, the life cycle of information in its broadest sense, logically-based game strategies, mechanisms for dual-use appropriation, end-to-end assessments, "routes to harm" in context, and multiple exposure pathways (Mueller,2021). It has been demonstrated that some DI technologies can lead to a redesign of the existing healthcare infrastructure. This might result in significant cost savings and make healthcare systems more sustainable. These DI technologies result in more person-centered healthcare by encouraging patients to become more active in their self-care, lowering the number of doctor's office and hospital visits, and the necessity for more extensive medical intervention. Medical devices and lifestyle/general-purpose apps will not be required to comply with the stricter definition, notified bodies, or clinical safety evidence. The stricter definition will not apply to lifestyle or general-purpose apps (Sheppard,2020). Laws and regulations can minimize complexity and uncertainty because they improve the predictability of the new technology. Despite the possibility of utilizing DI healthcare for its contribution to the healthcare architecture, the many regulatory processes intended to enhance patient confidence may be an obstacle to the success of this technology. There must be a balance between enabling patient trust and encouraging innovation (Sounderajah et al.,2021).

8. Discussion of results and future of DI in Healthcare

There exists noticeable uncertainty around using the phrase "disruptive innovation," even though it has become increasingly common in the healthcare business. Inaccurate identification might result in a lack of understanding regarding the characteristics and potential of an innovation. This can lead to a delay in translating its benefits into actual economic and health outcomes because we do not understand the probable barriers to adoption and how to overcome them. Finding examples of truly disruptive technologies in the health sector is hard. While there are suggestions to classify upgraded technology as DI, none of these technologies have been able to change the market in a big way. None have changed healthcare significantly or led to the creation of new and better ways to deliver healthcare. Using AI as an example, several innovations are complicated, providing the opposite of an expected disruption. Panch et al. (2019) found several hiccups in the quest for AI to usher positively in disruptive innovation. It was noted in the vein of predictive medicine through Hermansson and Kahan (2018) that overtraining an AI healthcare algorithm on Caucasian data produced poorer predictive care for non-Caucasian groups. This issue falls in line with the poor application of AI in other areas and is an example of how their use has far to go in suitability for routine, un-aided use without perpetuating negative or unhelpful feedback loops of care (Costanza-Chock et al. 2022; Hoffman,2021; Gebru,2020; Kong,2022). Many sources of electronic medical records systems hold incomplete connectedness, communication, and completeness, making it difficult for them to be used optimally (Patch et al.,2019). This could compound the quality of assistance being lent by AI assistants. With all applications of AI, and “innovations” in general, it can be their auditing, internally and externally, is important, especially when their impact remains new and poorly understood. Constanza-Chock et al., (2022) note a need for greater attention to AI audits performed by AI operators to reduce harm to stakeholders. Further, governments must be aware of and respond adequately to AI implementation in healthcare. The healthcare industry cannot function without contact between government and business through laws, professional standards, and administrative procedures. We need to be on the same page for significant disruption, and we are not there yet.

9. Concluding Remarks

Here,we provide a review and an analysis of a case study on the use of technology in healthcare services. We analyzed the approaches taken by the healthcare industry before the pandemic, during, and post-pandemic to recognize any impediment to genuine disruptive innovation. Major security lessons learned can be condensed in the understanding that efforts in implementing cybersecurity in hospital medical contexts are scattered throughout sites, are hampered by inefficiencies in organizational cultural absorption, and remain ultimately slow enough to be ahead of malicious actors. This relegates institutions to remaining reactive. This phenomenon

aligns with Sounderajah et al. (2021) that there has been little evident "disruptive innovation," only technological advancements. For "disruptive innovation" to exist, it must be accessible to everyone who is expected to utilize it. Further, there needs to be wider organizational and cultural technological literacy and agility to suitably adapt the components of such innovation. The aforementioned medical devices may become DI technology when extensive availability of the technology and ubiquitous base policy, legislation, and usage is the norm. Otherwise, we risk furthering hubris and wasting further resources on technical roads to nowhere.

Acknowledgements:

The authors would like to thank Jesse Finch for his contributions to editing.

References

Adler, A., Beal, J., Lancaster, M., Wyschogrod, D. (2021). Cyberbiosecurity and Public Health in the Age of COVID-19. In: Trump, B.D., Florin, MV., Perkins, E., Linkov, I. (eds) Emerging Threats of Synthetic Biology and Biotechnology. NATO Science for Peace and Security Series C: Environmental Security. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-2086-9_7

Agency outlines barriers, progress in adoption of nationwide electronic health information system | AAP News | American Academy of Pediatrics. (n.d.). Accessed October 18, 2022,https://publications.aap.org/aapnews/news/7392?autologincheck=redirected?nfToken=00000000-0000-0000-0000-000000000000

Anita Ramsetty, Cristin Adams, Impact of the digital divide in the age of COVID-19, Journal of the American Medical Informatics Association, Volume 27, Issue 7, July 2020, Pages 1147–1148, https://doi.org/10.1093/jamia/ocaa078

Bellucci, N. (2022). Disruptive Innovation and Technological Influences on Healthcare. *Journal of Radiology Nursing*.

Brooks, Anthony Lewis. "Disruptive innovation in healthcare and rehabilitation." In *Technologies of Inclusive Well-Being*, pp. 323-351. Springer, Berlin, Heidelberg, 2014.

Contact tracing: how physicians used it 500 years ago to control the bubonic plague. (n.d.). Accessed October 18, 2022https://theconversation.com/contact-tracing-how-physicians-used-it-500-years-ago-to-control-the-bubonic-plague-139248

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022, June). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571-1583).

Cybersecurity in Healthcare | HIMSS. (n.d.). Accessed October 18, 2022https://www.himss.org/resources/cybersecurity-healthcare

Eighth Broadband Progress Report | Federal Communications Commission. (n.d.). Accessed December 8, 2022, from https://www.fcc.gov/reports-research/reports/broadband-progress-reports/eighth-broadband-progress-report

Garrett, P., Brown, C. A., Hart-Hester, S., Hamadain, E., Dixon, C., Pierce, W., & Rudman, W. J. (2006). Identifying barriers to the adoption of new technology in rural hospitals: a case report. Perspectives in health information management, 3, 9.

Gebru, T. (2020). Race and gender. *The Oxford handbook of ethics of AI*, 251-269.

Haddara, M., & Staaby, A. (2018). RFID applications and adoptions in healthcare: a review on patient safety. *Procedia computer science*, *138*, 80-88.

Harris, D. A., Haskell, J., Cooper, E., Crouse, N., & Gardner, R. (2018). Estimating the association between burnout and electronic health record-related stress among advanced practice registered nurses. Applied Nursing Research, 43, 36-41.

He, Y., Aliyu, A., Evans, M., & Luo, C. (2021). Health Care Cybersecurity Challenges and Solutions Under the Climate of COVID-19: Scoping Review. Journal of Medical Internet Research, 23(4). https://doi.org/10.2196/21747

Hermansson, J., & Kahan, T. (2018). Systematic review of validity assessments of Framingham risk score results in health economic modelling of lipid-modifying therapies in Europe. *Pharmacoeconomics*, *36*(2), 205-213.

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. *New Media & Society*, *23*(12), 3539-3556.

Hwang, J., & Christensen, C. M. (2008). Disruptive innovation in health care delivery: a framework for business-model innovation. Health affairs, 27(5), 1329-1335.

Jalali, M. S., & Kaiser, J. P. (2018). Cybersecurity in hospitals: a systematic, organizational perspective. Journal of medical Internet research, 20(5), e10059.

Kilsdonk, E., Peute, L. W., & Jaspers, M. W. M. (2022). Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. International Journal of Medical Informatics, 98, 56–64. https://doi.org/10.1016/j.ijmedinf.2016.12.001

Kuehn, B. M. (2021). Despite Improvements, COVID-19's Health Care Disruptions Persist. JAMA, 325(23), 2335–2335. https://doi.org/10.1001/JAMA.2021.9134

Kong, Y. (2022, June). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 485-494).

Lang, M., Connolly, L. Y., Taylor, P., & Corner, P. J. (2021). The Evolving Menace of Ransomware: A Comparative Analysis of Pre-pandemic and Mid-pandemic Attacks. Digital Threats: Research and Practice. https://doi.org/10.1145/3558006

Lee, S. M., & Lee, D. (2021). Opportunities and challenges for contactless healthcare services in the post-COVID-19 Era. Technological Forecasting and Social Change, 167, 120712.

Liu, Z., Shi, Y., & Yang, B. (2022). Open Innovation in Times of Crisis: An Overview of the Healthcare Sector in Response to the COVID-19 Pandemic. Journal of Open Innovation: Technology, Market, and Complexity, 8(1), 21.

Maria K Sheppard, mHealth Apps: Disruptive Innovation, Regulation, and Trust—A Need for Balance, Medical Law Review, Volume 28, Issue 3, Summer 2020, Pages 549–572, https://doi.org/10.1093/medlaw/fwaa019

Moran, M. (2016). Barriers to Adopting HIT Still Substantial, APA Tells Health Subcommittee. Psychiatric News, 51(20), 1–1. https://doi.org/10.1176/APPI.PN.2016.10B10

Mueller S. Facing the 2020 pandemic: What does cyberbiosecurity want us to know to safeguard the future?. Biosaf Health. 2021;3(1):11-21. doi:10.1016/j.bsheal.2020.09.007

Adoption of electronic health records by hospital service type 2019-2021 (no date) HealthIT.gov. Available at: https://www.healthit.gov/data/quickstats/adoption-electronic-health-records-hospital-service-type-2019-2021 (Accessed: October 20, 2022).

Schulman, K., Vidal, A. & Ackerly, D. Personalized medicine and disruptive innovation: Implications for technology assessment. Genet Med 11, 577–581 (2009).https://doi.org/10.1097/GIM.0b013e3181ae0935

Sensmeier, Joyce E. MS, RN-BC, CPHIMS, FHIMSS, FAAN. Disruptive innovation and the changing face of healthcare. Nursing Management (Springhouse): November 2012 - Volume 43 - Issue 11 - p 13-14 doi: 10.1097/01.NUMA.0000421681.71712.86

Sera Whitelaw, Danielle M Pellegrini, Mamas A Mamas, Martin Cowie, Harriette G C Van Spall, Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review, European Heart Journal - Digital Health, Volume 2, Issue 1, March 2021, Pages 62–74,https://doi.org/10.1093/ehjdh/ztab005

Shaheen, M. Y. (2021). Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*.

Singh, S., Dhir, S., Das, V. M., & Sharma, A. (2020). Bibliometric overview of the Technological Forecasting and Social Change journal: Analysis from 1970 to 2018. Technological Forecasting and Social Change, 154, 119963.

Sounderajah V, Patel V, Varatharajan L, et al. Are disruptive innovations recognised in the healthcare literature? A systematic review. BMJ Innovations 2021;7:208-216.

The growing threat of ransomware attacks on hospitals | AAMC. (n.d.). Accessed September 20, 2022)https://www.aamc.org/news-insights/growing-threat-ransomware-attacks-hospitals

Thimbleby, H. (2013). Technology and the future of healthcare. Journal of public health research, 2(3), jphr-2013.

Types of Healthcare Information Systems - Adracare. (n.d.). Accessed December 15, 2022)https://adracare.com/2020/12/04/types-of-healthcare-information-systems/

WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. (n.d.). Accessed October 18, 2022https://covid19.who.int/

Why is the Healthcare Industry Slow to Adopt Technology? - WynHouse Software. (n.d.). Accessed September 20, 2022)https://wynsoftware.com/wynmill/healthcare-slow-to-adopt-technology/

Wynants, L., Smits, L. J., & Van Calster, B. (2020). Demystifying AI in healthcare. bmj, 370.

Yellowlees, P., Odor, A., Patrice, K., Parish, M. B., Nafiz, N., Iosif, A. M., & Hilty, D. (2011). Disruptive innovation: the future of healthcare?. Telemedicine and e-Health, 17(3), 231-234.

Securing Commercial Satellites for Military Operations: A Cybersecurity Supply Chain Framework

Courtney Fleming, Mark Reith, and Wayne Henry
Air Force Institute of Technology, Wright-Patterson Air Force Base, USA

courtney.fleming@afit.edu
mark.reith@afit.edu
wayne.henry@afit.edu

Abstract: The increased reliance on commercial satellites for military operations has made it essential for the Department of Defense (DoD) to adopt a supply chain framework to address cybersecurity threats in space. This paper presents a satellite supply chain framework, the Cybersecurity Supply Chain (CSSC) Framework, for the DoD in the evaluation and selection of commercial satellite contracts. The proposed strategy is informed by research on cybersecurity threats to commercial satellites, national security concerns, current DoD policy, and previous cybersecurity frameworks. This paper aims to provide a comprehensive approach for safeguarding commercial satellites used by the DoD and ensuring the security of their supporting components. Inspired by the National Institute of Standards and Technology (NIST) 800-171 requirements and the DoD's future Cybersecurity Maturity Model Certification (CMMC) process, the two-part framework significantly streamlines the NIST requirements to accommodate small businesses. It also extends key NIST requirements to commercial-off-the-shelf (COTS) suppliers. The CSSC Framework complements the CMMC certification process by addressing the need for cybersecurity requirements for all subcontractors supporting a commercial space asset. The framework incorporates a scoring process similar to CMMC scoring, granting points to a subcontractor for meeting the cybersecurity requirements outlined by the framework. In addition, the framework creates a space architecture overview that details the overall bid score and establishes a matrix based on individual requirements. This model and matrix allow DoD acquisition personnel to closely analyze each contract bid, comparing the subcontractor's strengths and weaknesses to other bidders. The CSSC Framework will allow the DoD to apply NIST standards to subcontractors who do not meet the requirements for CMMC certification.

Keywords: Space Systems, Commercial Satellites, Supply Chain, Space Cybersecurity, Commercial-Off-The-Shelf

1. Introduction

In recent years, the commercial space industry has grown rapidly, making it easier and cheaper for businesses and academic institutions to deploy satellites in low-earth orbit. A domain that was once dominated by governments around the globe is now accessible to almost any type of organization. However, the market demand for these quickly deployable systems highlights a growing cybersecurity concern. Commercial satellite systems may require multiple vendors within a complex supply chain, many of which do not follow centralized, standard cybersecurity requirements. The DoD recognizes the importance of addressing cybersecurity in space systems and acknowledges the potential national security threats posed by supply chain risks. However, there is currently no framework outlining how the DoD can secure the space cybersecurity supply chain.

This paper proposes a Cybersecurity Supply Chain (CSSC) framework outlining the requirements that contractors and subcontractors within the space supply chain should follow in order to bid for a DoD contract. Before presenting the CSSC Framework, the paper first discusses cybersecurity threats to commercial satellites, including a recent state-sponsored attack on a US commercial satellite. This paper also examines recent research by the National Institute of Standards and Technology (NIST), the Aerospace corporation, and the MITRE corporation, as well as current DoD policy related to space, cybersecurity, and the supply chain. In addition, the paper briefly describes a general satellite architecture and uses this architecture to propose the two-part framework.

2. Cybersecurity threats to commercial satellites

Historically, commercial satellite developers have not prioritized cybersecurity considerations when creating space systems, similar to the challenges at the inception of the Internet. However, as the space domain's attack surface expands with the increase of satellites in orbit, experts are concerned with cybersecurity threats posed to commercial satellites, whether privately owned or used by the government (House of Representatives, 2022). For example, Falco (2019) highlights that satellites can be a single point of failure for entire industries. Additionally, the supply chain for space systems can be long and involve many stakeholders. The use of

commercial-off-the-shelf (COTS) software within these systems is also a concern. Furthermore, there is a lack of regulation and a need for space system cybersecurity standards when evaluating space products (Falco et al., 2022). These factors make satellites and their components vulnerable targets for adversaries, including state and non-state actors, to exploit. This is especially concerning as these single points of failure can affect communication, navigation, and imagery for millions of people.

The use of COTS software and hardware components presents unprecedented challenges for commercial satellites, as the presence of these assets in space has diversified from government control to CubeSats deployed by many businesses. COTS components allow companies and organizations to launch satellites at a low cost, but they also have the potential for vulnerable components due to fast production and little security consideration. The use of COTS components allows attackers to easily access the code to explore for vulnerabilities, which was previously more difficult when satellites used proprietary software (Nussbaum and Berg, 2020). Despite swift patches to discovered vulnerabilities, significant challenges impede patching to space systems in orbit, and many remain unpatched. Even though publicized attacks on in-orbit satellites are limited, cyber upkeep on ground-based components and strong network security for components in space is still a need (Nussbaum and Berg, 2020).

Similar to large technology companies, space systems face many of the same challenges as those in terrestrial cyberspace. One study analyzed satellite incidents between 1977 and 2019, categorizing them based on both the architecture and industry targets. Most attacks targeted the ground segment of government satellite architectures (Manulis et al., 2021). Manulis et al. (2021) further assessed that most of these attacks were conducted for state espionage, political gain, and corporate espionage, and employed techniques such as jamming, computer network exploitation, hijacking, and phishing. Over the last decade, the number of satellites in orbit increased exponentially, and the amount of satellite incidents follow this trend.

3. Recent example: state attack on US commercial satellite

As the DoD expresses its intent to incorporate commercial satellites into military operations, it is important to analyze cyber attacks aimed at these dual-use satellites to understand the threats imposed by near-peer adversaries. The most recent example is the Russian cyber attack in February 2022 on satellites of US company, ViaSat, which supported Ukrainian telecommunications during the Russia-Ukraine conflict. Boschetti et al. (2022) conducted an in-depth review of this cyber attack, highlighting its significance as the only publicly known attack against a commercial space system used by a government. While the details about the attack are unknown, Boschetti et al. (2022) present a hypothetical attack cycle based on statements made by ViaSat. In 2022, attackers exploited a vulnerability discovered years prior within the VPN “Fortigate”. Boschetti et al. (2022) further elaborate that the ViaSat attack began with the attacker using this unpatched VPN to access ViaSat ground segments, which were maintained by a subsidiary of another satellite company, to finally access ViaSat’s modem management interface. From there, the attackers uploaded wiper malware and disabled communications for thousands of users.

This attack demonstrates the need for vetting contractors to ensure their third-party provided components are secure, that the contractors use a zero-trust architecture, and that contractors practice quality vulnerability checks and patch management (Boschetti et al., 2022).

4. National security concerns

In July 2022, experts representing the MITRE Corporation, NIST, and the Aerospace Corporation testified in front of the Subcommittee on Space and Aeronautics, under the House Science, Space, and Technology Committee regarding satellite cyber threats to the United States. Their testimony expressed concern for cybersecurity attacks on satellites causing a collision with another satellite in-orbit. Specific examples of attacks affecting satellite operations include using an injection of malicious code to spoof sensor data, corrupt sensor systems, or unauthorized commands for guidance and control (House of Representatives, 2022). Their testimonies further outlined the criticality behind securing space systems. The Defense Intelligence Agency (DIA) assessed that foreign adversaries plan to target space systems through cyber operations. Current security limitations include a lack of a widely adopted technical standard for commercial space assets and a lack of information sharing within the research and development sector for space technology.

5. Recent research efforts

Researchers now acknowledge the threats to space cybersecurity and propose frameworks to address these threats. The Aerospace Corporation led the initiative to create the Space Attack Research and Tactic Analysis (SPARTA), a matrix based upon the MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) framework, a knowledge base of tactics, techniques, and procedures (TTPs) demonstrated by adversaries to launch cyber attacks. The MITRE ATT&CK model lists and explains techniques for each stage of the cyber attack including reconnaissance, initial access, privilege escalation, command and control, and impact. It also lists mitigation measures and recommendations to combat the adversary's TTPs (MITRE Corporation, 2022). SPARTA mirrors MITRE's matrix and applies it to space cyber attacks, listing TTPs specific to space to gather information from the space and cybersecurity communities and fully address and share adversarial TTPs (SPARTA, 2022). Another characteristic unique to SPARTA is it incorporates references from NIST that outlines recommended countermeasure controls, which experts found most relevant to space. Both frameworks acknowledge that supply chain compromise is a tactic used by adversaries to gain initial access in order to aid a cyber attack.

Additional research efforts to address the threats in the space domain include a recent collaboration between NIST and the MITRE corporation to provide an existing cybersecurity framework for commercial space operators to apply to their systems. Scholl and Suloway (2022) outline their work as supporting Space Policy Directive 5 (SPD-5) which directs government agencies to define cybersecurity norms for those operating in the space domain. Their work describes the threats to commercial satellite systems and the Cybersecurity Framework created from Executive Order 13636, Improving Critical Infrastructure Cybersecurity. The functions within the framework are "identify, protect, detect, respond, and recover." Scholl and Suloway (2022) provide a matrix of subcategories supporting the five functions, which detail specific actions commercial satellite operators can take or self-evaluate to ensure their infrastructures are secure. NIST and the MITRE Corporation are still refining this product, and the actions listed within this research are only recommendations at this point.

6. Current DoD policy

Little information is available regarding the DoD's implementation of a space-specific cybersecurity framework to address supply chain threats. However, in December 2021 the DoD released the Cybersecurity Maturity Model Certification (CMMC) Version 2.0 Overview to provide an updated framework for protecting federal contract information and controlled unclassified information within the supply chain. The CMMC model measures the implementation of cybersecurity requirements based upon the Federal Acquisition Regulation (FAR) and NIST standards and divides these measurements into three levels: Level 1 consists of 17 "foundational" practices recommending an annual self-assessment of these practices. Level 2 consists of 110 "advanced" practices recommending triennial third-party assessments. The incomplete level 3 is intended to contain "expert" practices from NIST SP 800-172 for triennial government assessments (CMMC Model Overview, 2022). The CMMC model consists of 14 domains including, but not limited to, access control, awareness/training, identification/authentication, and security assessment. An example of practices by level under these domains is insider threat training, meeting the NIST 800-171 Rev 2 3.2.3 requirement, which is classified as a Level 2 practice under awareness/training (CMMC Model Overview, 2022). While contractors are already required to maintain cybersecurity standards outlined by NIST 800-171, COTS component suppliers are exempt from the CMMC certification requirement (General Services Administration, 2022). A survey of multiple businesses following the previous CMMC Version 1.0 model revealed that small businesses were more likely to struggle with the CMMC compliance (Strohmier et al., 2022).

Additionally, the DoD enforces policy regarding open-source/COTS software, which is useful when considering satellite software, such as flight software. The Office of the DoD Chief Information Officer (CIO) published answers to frequently asked questions relating to open-source software. The office asserts that it encourages the use of open-source and COTS software as it is generally inclined to be more secure due to more developers and reviewers being able to identify vulnerabilities, discrediting the "security by obscurity" theory (Office of DoD CIO, 2021). It frequently references a study from 2003 produced by MITRE, analyzing the use of open-source software in the DoD, which concludes that open-source software played a critical role in the DoD. Without it, there would be negative consequences, limiting the DoD from protecting its network (MITRE Corporation, 2003). For security, this study recommends that the DoD develop a list of open-source software that is "generally recognized as safe," but the DoD has not developed such a list. According to the Office of DoD CIO, the DoD requires commercial software to come with a warranty or source code to ensure the supplier or government can maintain the software.

For general DoD supply chain threats, the DoD published an action plan in response to Executive Order 14017 to address defense supply chains (US DoD, 2022). Its focus areas as most pressing to national security include kinetic capabilities, energy storage and batteries, castings and forgings, and microelectronics and include "cyber posture" as a strategic enabler, encompassing industrial security, counterintelligence, and cybersecurity. The document briefly mentions space as a supply chain concern, referencing spacecrafts as a platform containing cast and forged parts, but it lacks cybersecurity considerations.

7. Commercial satellite components/architecture

This section presents a brief overview of commercial satellite general components and architecture to fully understand the diversity of potential suppliers that support the architecture. Willis et al. (2017) describe a notional satellite as being similar to an industrial control system, made up of systems within systems connected via buses. The main elements within the architecture are propulsion, thermal, power, attitude, and data handling subsystems. Similarly, Scholl and Suloway (2022) provide an overall, general architecture of satellite operations which divides into three segments: the space segment, the ground segment, and the user segment. The space segment consists of a link sub-segment, internal satellite cybersecurity sub-segment, and a satellite-to-satellite communication sub-segment. The ground segment is terrestrial operations conducted by humans, which can be co-located with launch facilities, separate from these facilities, or outsourced altogether. The user segment describes the endpoints for the consumers such as GPS receivers, telecommunication equipment, and aircrafts. The following proposed framework omits the user segment and uses the space and ground segments described by Scholl and Suloway (2022) as a reference.

8. Proposed framework

This section proposes a notional framework for acquisitions to understand the contractor's supply chain from launch to orbit and outlines cybersecurity requirements that each of the contractor's suppliers must meet to bid for a DoD contract. Rather than relying on the CMMC model to ensure cybersecurity compliance, the CSSC Framework uniquely compliments the CMMC model by extending cybersecurity requirements to COTS suppliers. The CSSC Framework also applies to any contractor that supports the commercial satellite regardless of the contractor's likelihood to store or handle controlled unclassified information. The framework's supporting cybersecurity conditions are streamlined from the CMMC model, reducing the domains from 14 to five, to overcome the complexity of a long supply chain with multiple COTS suppliers. The five domains are network management, identification/authentication, system management, general security, and incident prevention/response. The new domains combine and summarize the requirements outlined in the CMMC model while meeting the key provisions of NIST 800-171. Also similar to the CMMC certification process, the framework implements a scoring system which provides an overall "bid score", consisting of a summation of the sub-scores from the five requirement categories. Like the CMMC certification process, the CSSC Framework requires a self-assessment and a third-party assessment before a contract bid submission. The following subsections explain the criteria for each of the five domains.

8.1 Network management

This domain focuses on the subcontractor's ability to properly enact security measures within all networks that the subcontractor owns or any other networks interfacing with the subcontractor's networks. Some of the concerns within the category include remote access, open ports, and secure protocols, which can lead to vulnerabilities if these network features are not properly configured. The CMMC Model Overview (2022) inspired this category. The framework summarizes requirements affecting the network and includes lessons learned from the ViaSat attack (Boschetti et al., 2022), resulting in four requirements. First, a company needs a zero-trust network architecture and must demonstrate how it verifies and controls connections to external networks as necessary. A third-party assessment may evaluate the strength of the zero-trust architecture. A supplier must also have a Virtual Private Network (VPN) connection in place, from a reputable provider, for wireless, mobile, and remote access that also prevents split tunneling. Network managers must also disable nonessential ports, protocols, and services, implementing a "least functionality necessary" approach to the networks. Finally, company networks need a policy that denies network communications traffic by default and allows traffic by exception.

8.2 Identification and authentication

This domain evaluates how well the subcontractor enforces identification and authentication practices for anyone attempting to access its systems or networks. The domain combines requirements from the "Identification and Authentication" and "Audit and Accountability" domains from the CMMC Model Overview (2022), which results in four requirements. First, the company needs a multifactor authentication technique. One effective example is issuing physical dongles to users that generate a random key to enter for system access in addition to using a smart card to plug into a terminal. For password implementation as part of multifactor authentication, the system must require a level of complexity and periodical changes with no repeating passwords while maintaining practical usability. Additionally, the company establishes a limit on unsuccessful login attempts, and the system must keep a log of unsuccessful login attempts. All systems and websites maintained by the company may only store and transmit cryptographically protected passwords, meaning the system must have an effective, non-proprietary hashing algorithm that obscures user passwords. Lastly, any systems and websites shall obscure stored authentication information. At a minimum, usernames and passwords cannot be stored in clear text.

8.3 System management

This domain focuses on the subcontractor's ability to properly enact security measures within all systems that the subcontractor owns. Adopted from multiple domains of the CMMC Model Overview (2022), primarily "System and Communications Protection" and "System and Information Integrity", there are seven requirements aligned with NIST 800-171. The company demonstrates the limitation of information system access based on absolute "need-to-know" or "need-to-access" principle of authorized users. Also, privacy and security notices/banners must be present during a user's initial system access to assist in incident response. All systems need automatic session locks and termination after a reasonable period. Any information technology products require established security configuration settings within the organizational system. System managers must configure only essential capabilities for organizational systems, adopting a "least functionality necessary" approach. Similar to the network management policy requirement, systems must enforce a "deny-all, permit by-exception" policy for the installation and execution of software within the system. Lastly, the subcontractor must demonstrate routine maintenance procedures and file scanning for systems periodically.

8.4 General security

This domain encompasses requirements that implement overall cyber-secure policies and practices. This includes training, media handling, physical security, and hardware screening, summarizing multiple domains from the CMMC Model Overview (2022). The six requirements include a company's implementation of routine training and awareness for role-based risk/responsibilities, insider threats, elicitation, phishing attacks, and media handling. The company also needs a policy for media handling, marking, accountability, and disposal. Within facilities, employees are required to easily identify authorized personnel to screen individuals on the premise. In the event an employee is terminated or re-assigned to another position, security managers re-evaluate and/or remove all identification characteristics, permissions, and accesses to systems following duty changes. Additionally, the company requires general facility security practices to include escort procedures, access logs, and additional barriers to screen for authorized personnel continuously. Finally, as applicable for critical hardware, the supplier/design team must ensure there are quality assurance inspections through design and use of hardware to identify defects or counterfeit components.

8.5 Incident prevention and response

This domain requires subcontractors to ensure any actions taken on networks, systems, and individual accounts are logged and analyzed while also ensuring these actions can be correctly attributed to the suspect user account. The seven requirements primarily summarize the "Incident Response" domain from the CMMC Model Overview (2022). System and network managers must conduct routine analysis and monitoring of audit logs for suspicious and unauthorized use. System capabilities require the synchronization of internal system clocks for audit records. Networks and systems must trace any actions back to a unique user. The company needs an established policy outlining detection, analysis, containment, and recovery incident-handling capabilities if there is suspicion that an intrusion occurred within the organization's systems. The contractor requires an incident reporting plan to notify all affected stakeholders and appropriate officials. Company cybersecurity professionals must conduct internal vulnerability scans and search for vulnerabilities published in open-source forums that

affect the organization. Lastly, software managers shall employ reputable static and analysis tools to look for vulnerabilities within the organization's software.

The first part of the CSSC Framework presents a way for a major company with multiple subcontractors under them to clearly outline the supply chain involvement so the DoD can properly vet them. This part also lists the suppliers' overall bid scores and affiliated foreign countries involved in the suppliers' operations. The foreign country affiliation requirement allows the DoD to ensure components do not originate from near-peer adversaries, and that an adversary does not influence a company's work with the DoD. Figure 1 depicts a supplier overview that the contractor would provide to the DoD, based on the satellite architecture described by Suloway and Scholl (2022). The framework divides the suppliers into two categories: the ground segment and the space segment. The ground segment is further divided into three subcategories: mission operations center, payload control center, and bus interfaces. Rather than divide the mission operations buses and payload buses into their own subcategories, all buses and links are compiled into one subcategory for brevity. Each subcategory lists the supporting supplier. For example, in Figure 1 Supplier C and Supplier D provide hardware and software that support the payload control center. In this scenario, any suppliers who provide communication capabilities from the ground segment to the space segment are classified under "bus interfaces."

Similarly, the space segment is divided into three subcategories: hardware components, embedded software, and internal communication. Each subcategory lists the supporting supplier. In Figure 1, for example, Supplier H and Supplier I provide software for the satellite's embedded systems. Any bus interfaces within the satellite are classified under "internal communication." The figure depicts an example of the number of suppliers supporting the overall commercial satellite architecture, and this number may vary. Additionally, a supplier may provide more than one function. For instance, a supplier supporting the mission operations center could also supply a component supporting the payload control center.

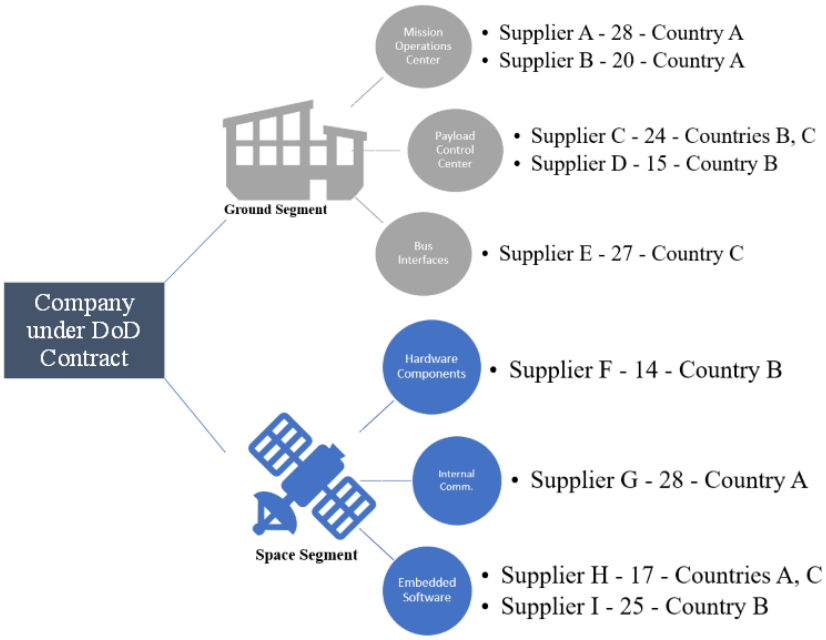


Figure 1: An overall depiction of each supplier providing services to the DoD contractor. This overview divides into the ground and space segments, listing the main capabilities of each segment. Each supplier providing the capability has an associated overall bid score and affiliated foreign countries providing support to the company.

Each supplier in every subcategory has an overall bid score. The scores are based on each requirement listed within the five domains supporting the underlying framework. If a supplier meets a requirement, it receives one point. Suppliers may not receive any points for partially satisfied requirements. The maximum bid score a supplier may receive is 28 points. In Figure 1, Suppliers A and G have a bid score of 28, meaning it has satisfied all cybersecurity requirements within each domain. Every supplier also has its affiliated foreign countries, and in Figure 1, countries B and C provide operational or financial support to Supplier C.

The second part of the CSSC Framework outlines a matrix for DoD acquisitions personnel to evaluate the supplier's strengths and weaknesses within each domain. The matrix is color-coded green, yellow, and red to symbolize domain requirement compliance. If the supplier meets all requirements within a domain, its cell is green. If the supplier meets over 50% of the requirements within a domain, its cell is yellow. For failing to meet 50% or more of the requirements, a supplier's cell is red. Figure 2 represents a matrix to accompany the example provided in Figure 1. Since Suppliers A and G have a perfect bid score, all their cells are green. The red cells in the "Hardware Components" row indicate Supplier F failed to meet more than 50% of the requirements in network management, identification/authentication, and system management. However, the yellow cells in the same row show that Supplier F met between 50% and 100% of the requirements in general security and incident prevention/response. Through the example used in Figure 2, an acquisition officer may deem a contractor is unsuitable considering the overall incident prevention/response requirement deficiencies.

	Network Management	Identification and Authentication	System Management	General Security	Incident Prevention and Response
Mission Operations Center	Supplier A	Supplier A	Supplier A	Supplier A	Supplier A
	Supplier B	Supplier B	Supplier B	Supplier B	Supplier B
Payload Control Center	Supplier C	Supplier C	Supplier C	Supplier C	Supplier C
	Supplier D	Supplier D	Supplier D	Supplier D	Supplier D
Bus Interfaces	Supplier E	Supplier E	Supplier E	Supplier E	Supplier E
Hardware Components	Supplier F	Supplier F	Supplier F	Supplier F	Supplier F
Internal Communication	Supplier G	Supplier G	Supplier G	Supplier G	Supplier G
Embedded Software	Supplier H	Supplier H	Supplier H	Supplier H	Supplier H
	Supplier I	Supplier I	Supplier I	Supplier I	Supplier I

Figure 2: A matrix demonstrating the strengths and weaknesses of each supplier's compliance for each domain. The matrix is color-coded to reflect the level of compliance for the requirements outlined in the domains. This provides a depiction to acquisition personnel of the cybersecurity viability for the space asset.

The benefits of the two-part framework include the enforcement of the NIST 800-171 standards for any contractor or subcontractor, including COTS suppliers. Additionally, the extension of these standards to all contractors, regardless of controlled unclassified information handling, mandates strong business practices and cybersecurity principles if they want to work with the DoD. The CSSC Framework also reduces the 127 requirements outlined in Levels 1 and 2 of the CMMC Version 2.0 model to 28 requirements. This alleviates the budgeting and planning overhead for small businesses who already struggle with the CMMC certification process. Lastly, the DoD acquisitions personnel reviewing the contract bid may use the framework for a detailed visual representation of the overall space architecture. DoD experts can compare the strength of each bidder's space and ground segments and its components, identify overall domain deficiencies, and understand the foreign countries tied to each subcontractor for supply chain threat management.

9. Limitations and future work

While there are several ways to evaluate the effectiveness of the CSSC Framework, open-source research is scarce since commercial satellite websites advertising their product do not provide a full list of their suppliers. Additionally, it is difficult to evaluate how well these suppliers implement cybersecurity practices unless there is publicly available information that reports negative publicity such as intrusions or data theft. Future research may find practicality in working directly with contractors and their suppliers/subcontractors to evaluate this proposed supply chain framework. However, there may be challenges in collaborating with suppliers due to their reluctance to publicly disclose possible vulnerabilities in their information technology infrastructure. As a result, research on this topic may be limited to distribution within the DoD only.

Among the several ways to evaluate the framework effectiveness is addressing the balance between enforced cybersecurity principles and the supply chain's ability and willingness to propose a bid that complies with these standards. One question to consider is whether the CSSC Framework requirements are too loose or unrealistic for subcontractors to follow, especially for COTS suppliers. A survey of commercial satellite COTS suppliers would help answer this question. Another consideration is the implementation of a DoD-led inspection to evaluate compliance. For example, the CMMC Model Overview (2022) enforces annual or tri-annual inspections by

government or third-party entities. A concern for a complex supply chain is having adequate DoD manpower to inspect every supplier or if more inspections are needed due to commercial satellites' mission-essential capabilities. An additional factor to consider is a comparison between this streamlined framework and the CMMC 2.0 model. While the CSSC Framework is intended to compliment the CMMC 2.0 certification process, future work could analyze if the framework eliminates critical requirements from the CMMC 2.0 model.

10. Conclusion

The space domain was previously restricted to state governments, and only a few decades ago, there were not as many deployed satellites in low-earth orbit. Commercial nanosatellites now clutter low-earth orbit, and it only increases the attack surface for cyberattacks against space systems. The DoD wants to use commercial satellites to facilitate its intelligence and detection capabilities. However, a framework and standard for cybersecurity within the space supply chain is missing to ensure the DoD can operate safely in space as it relies on commercial satellite operators. This paper proposes a two-part framework that enables transparency within the commercial space supply chain and outlines requirements that contractors and suppliers must meet to compete for a contract. The paper also recommends ideas and challenges for future researchers to evaluate the CSSC Framework and compare it to an existing DoD cybersecurity framework.

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

Aerospace Corporation. (2022). *Software Dependencies & Development Tools, Technique IA-0001.01 | SPARTA.* [online] Available at: https://sparta.aerospace.org/technique/IA-0001/01/.

Boschetti, N., Gordon, N.G. and Falco, G. (2022). Space Cybersecurity Lessons Learned from The ViaSat Cyberattack. In ASCEND 2022 (p. 4380).

Falco, G. (2019). Cybersecurity principles for space systems. Journal of Aerospace Information Systems, 16(2), pp.61-70.

Falco, G., Henry, W., Aliberti, M., Bailey, B., Bailly, M., Bonnart, S., Boschetti, N., Bottarelli, M., Byerly, A., Brule, J. and Carlo, A. (2022). An International Technical Standard for Commercial Space System Cybersecurity-A Call to Action. In ASCEND 2022 (p. 4302).

General Services Administration. (2022). 252.204-7021 Cybersecurity Maturity Model Certification Requirements. | Acquisition.GOV. [online] Available at: https://www.acquisition.gov/dfars/252.204-7021-cybersecuritymaturity-model-certification-requirements.

House of Representatives. (2022). *Exploring Cyber Space: Cybersecurity Issues for Civil and Commercial space systems: House Committee on Science, Space and Technology*. [online] Available at: https://science.house.gov/hearings/exploring-cyber-space-cybersecurity-issues-for-civil-and-commercial-space-systems.

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.

MITRE Corporation. (2022). *MITRE ATT&CK*. [online] Available at: https://attack.mitre.org/.

MITRE Corporation. (2003). Use of Free and Open-Source Software (FOSS) in the U.S. Department of Defense. [online] Available at: https://dodcio.defense.gov/Portals/0/Documents/FOSS/dodfoss_pdf.

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.

Office of the DoD Chief Information Officer. (2021). *Open Source Software FAQ.* [online] Available at: https://dodcio.defense.gov/open-source-software-faq/#q-is-oss-commercial-software-is-it-cots.

Scholl, M. and Suloway, T. (2022). Introduction to Cybersecurity for Commercial Satellite Operations (2nd Draft), *National Institute of Standards and Technology*.

Strohmier, H., Stoker, G., Vanajakumari, M., Clark, U., Cummings, J. and Modaresnezhad, M. (2022). Cybersecurity Maturity Model Certification Initial Impact on the Defense Industrial Base. JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH.

United States Department of Defense (DoD). (2022). *Cybersecurity Maturity Model Certification (CMMC) Model Overview*. [online] Available at: https://www.acq.osd.mil/cmmc/docs/ModelOverview_V2.0_FINAL2_20211202_508.pdf.

United States Department of Defense (DoD). (2022). *Securing Defense-Critical Supply Chains.* [online] Available at: https://media.defense.gov/2022/Feb/24/2002944158/-1/-1/1/DOD-EO-14017-REPORT-SECURING-DEFENSE-CRITICAL-SUPPLY-CHAINS.PDF.

Willis, J.M., Mills, R.F., Mailloux, L.O. and Graham, S.R. (2017). Considerations for secure and resilient satellite architectures. In 2017 International Conference on Cyber Conflict (CyCon US) (pp. 16-22). IEEE.

Predictors of Human Efficiency in Radar Detection Tasks

Elizabeth L. Fox[1], Arielle L. Stephenson[2], Christopher A. Stevens[1], Gregory Bowers[2]
[1]Air Force Research Laboratory, Wright Patterson Air Force Base, OH, USA
[2]Ball Aerospace, Dayton, OH, USA

elizabeth.fox.9@us.af.mil
arielle.stephenson@ballaerospace.com
christopher.stevens.28@us.af.mil
gregory.bowers@ballaerospace.com

Abstract: Aegis operators simultaneously locate and monitor the activity of several hostile targets, intervening and alerting their team when appropriate. Utilizing the Aegis Combat System, operators find, track, and respond to dynamic targets on a radar screen. The demand that operators undergo is often high, inevitably causing strain on cognitive functions and detriments to performance. We applied model-based measures, Cost and Multitasking Throughput, to quantify the influence of external factors on processing efficiency in radar task(s). We captured the influence of three experimental manipulations, each of three levels, on human efficiency to track the location of hostiles and/or detect brief radar interference. We collected participants' performance to complete a multiple object tracking (MOT) task and an electronic attack detection task (EA) using a radar display. A factorial manipulation of conditions comprised changes to task(s) (EA, MOT, or both), the number of targets to track (2, 4, or 6) and the presence or absence of distractors, deemed 'friendlies' (between 500-1000 total tracks). Our novel individual- and model-based approach provided quantitative estimates of human efficiency. We compared the observed variation in efficiency among predictors including target quantity, visual load, and the presence of one or two interrelated tasks. Through quantifying the relationship of these variables to radar detection tasks, we discuss implications of our findings and provide a framework to examine how system designers may develop tools to alleviate observed cognitive demands and/or counter potential threats of electronic attacks in radar detection and tracking tasks.

Keywords: Radar, Electronic warfare, Human efficiency, Dual-tasking

1. Introduction

Radar operators are responsible for the detection and tracking of entities that pose a threat to military operations or civilian lives. To make accurate and timely decisions, operators rely on the information portrayed through integrated and dynamic visual displays that comprise the input from several powerful electronic radar systems. An Aegis Combat System (ACS) is a platform equipped with protective software to simultaneously track hundreds of entities, continuously updating a visual display for operators to monitor threats and make mission critical decisions in an efficient manner (see Bath, 2020, for a recent overview). An Aegis system on board vessels can provide tactical data such as automatic target labeling and display this information to operators with the aim to facilitate efficient and accurate crisis management. Advanced Aegis technology may fuse input from multiple radars regarding attributes such as entity position, speed, and size to assign and convey meaningful labels to operators by-use-of symbology or text. Effective user interface design can lead to faster identification of potential adversaries, increase the probability of mission success, and reduce casualties of civilian and military personnel.

Importantly, the high-risk and sensitive nature of the information portrayed by an ACS necessarily requires a human operator to posture the appropriate course of action (versus an automated response), following their interpretation and assessment of entities using a graphic user interface (GUI). This heightens the priority for GUI designers of an Aegis system to understand and account for limitations and strengths of human perception and cognition, especially when attempting to make sense of complex and multifaceted dynamic visualizations. For instance, an Aegis system may have no problem reshuffling the use of target tracking numbers as it updates incoming data, but a human operator may commit a unique identifier to a single target in memory. This failure to share a common cognitive representation of entities may lead to false threat identification. Indeed, in 1988 USS Vincennes mistakenly shot down Iran Air Flight 655 killing 290 civilians due to poor human-system integration, lamentably earning recognition as one of the worst outcomes of poorly executed user-interfaces (Pogue, 2016). In addition, the under-evaluated consequences from cuts to training or system maintenance can leave crew members and ACS unequipped for battle readiness (Fisher and Kingma, 2001). As a result, military costs may be reduced but the gap widens between humans and complex radar displays they leverage to detect and act on threats and plan crucial mission operations.

Little research has aimed to investigate how and why human performance changes in realistic threat detection scenarios. Nonetheless, this work is crucial to build effective GUI and intelligent Aegis software and foster

efficient human-system integration. The aim of our work described in this paper was to capture and quantify variation in human efficiency to perform tasks similar to that of an Aegis operator while undergoing systematic contextual and cognitive changes relatable to those that may occur in operational environments. We measured human efficiency to follow various numbers of tracks labeled as hostiles (2, 4, or 6) as they dynamically travel across four geographical areas of interest (AoI), occasionally changing in their heading direction. Participants monitored these entities in isolation or in the presence of several hundred friendly tracks. Friendly tracks (i.e., ships, aircraft, drones, etc.) were perceptually distinct from, but closely resembled, hostiles.

In addition to the detection and tracking of hostiles, realistic operations require Aegis operators to be vigilant against electronic attacks (EA). Electronic attacks, in part, make-up electronic warfare (EW). Electronic warfare is divided into three categories, all of which aim to "protect friendly operations and deny adversarial operations within the electromagnetic (EM) spectrum throughout the operational environment" (Joint Chiefs of Staff [JCS], 2013). Electronic attacks involve the use of energy (e.g., EM) to intentionally target an enemy and adversely impact their military capabilities (JCS, 2013). Algorithms to both deploy and counter EA (called electronic protection or EP) are constantly (and necessarily so) developed and revised to account for endless technological advances (e.g., Park et al., 2018; see Spezio, 2002 for a review). An EA may affect one or multiple entities, may be one or multiple modalities (i.e., chaff, cyber, jammer), and may endure for various lengths of time. The attempted EA and its effect on ACS information will depend on the attackers' goals and the (counter-)technology deployed. However, the research field lacks psychological understanding for how the type or timing of EA may differentially influence how operators detect and respond to it, let alone how the mere *possibility* of an attack will influence how an operator performs and makes decisions about entities represented on a GUI – in this work, we focus on the latter. We argue it is pertinent to consider EA in the broader context, that is, embedded as an additional demand among multiple tasks which compete for operators' limited attention and perceptual resources (Kahneman, 1973; Norman and Bobrow, 1974). Monitoring for the presence of EA increases the mental load of an operator and heightens the likelihood of detrimental interruptions and failures to detect and track adversaries.

As mentioned above, how an EA influences the entity (entities) that are represented on the GUI and shown to the operator depends on both the attackers' goals and the EP in place to counter or partially (fully) block the attack. In this work, we instantiated EA through the disappearance of a single, threatening target for one to four seconds without warning – a simplified case of radar *jamming.* It is reported that jamming is one type of EA that may cause an ACS to temporarily lose a target (Blair et al., 1998). Reacquisition and labeling of a lost entity requires sophisticated decision making processing, and current intelligent radar systems still rely on human intervention to appropriately identify and respond to radar jamming (e.g., the lost target was an enemy threat) (Gurbuz et al., 2019). We asked participants to indicate when an EA began (i.e., when a threatening entity disappeared) and when it ended (when the entity reappeared). In addition to manipulations to the number of hostiles (2, 4, or 6) and presence of friendlies described above, we systematically introduced instances of EA on a single to-be-tracked (marked as threatening) entity and compared human efficiency to 1) track hostiles and 2) detect instances of EA in an isolated versus dual-task context.

2. Method

We investigated how the number of targets (2, 4 or 6), the presence (absence) of friendlies, and the nature of the task(s) demands (tracking target location, detecting EA, or both) influences participants' efficiency to make decisions and correctly respond to events – our independent variables, a 3 x 2 x 2 within-subjects design. We computed individuals' Multitasking Throughput (MT; Fox, et al., 2021), a performance-based model of efficiency, to quantify systematic changes among our independent variables of interest – our dependent variable. We hypothesized that efficiency would 1) decrease as the number of targets to track increased, 2) decrease with the presence of friendlies, 3) decrease when performing both tasks, compared to each in isolation.

Our method and unique modeling framework allowed us to equate the intersection between each of our independent variables and efficiency at the individual participant level. For instance, the degree of efficiency deficit observed when tracking 6 targets when no friendlies are present may be equivalent to the deficit observed when tracking 4 targets when friendlies are present and/or the deficit observed when responsible to track the location of 2 targets when friendlies are present and EAs may occur. By identifying such relationships of these variables, our results have theoretical and practical implications: we provide a quantitative mapping between three manipulations of cognitive workload (number of targets, presence of visual distractors, and dual-tasking) that may inform the design of cognitive radar technology to help mitigate or alleviate the workload of

operators and help guide the focus to build effective countermeasures of potential threats (e.g., the design of EP software).

2.1 Participants

An a priori power analysis (G*Power 3.1.9.6) indicated we should collect data from 28 participants for an appropriately powered within-subjects, three factor, repeated measure analysis of variance (ANOVA). To obtain this estimate, we assumed a medium effect size (f = .25), power (1-ß = .80), significance level (α = .05), and correlation (r = .50), with a correction for nonsphericity (∈ = 1.0) (see Kang, 2021 for a full description and guide for using G*Power). We recruited 29 participants from a medium-sized, midwestern university via paper flyers and an online announcement or from a long-term subject panel (Wright-Patterson Air Force Base) and compensated with either visa gift cards ($20/hour + $10 study completion bonus) or through their normal wage, respectively. Participants' (9 = Male, 19= Female, 1 = Other) ages ranged from 19-57 (M = 26.14, SD = 8.44). All participants self-reported normal or corrected to normal hearing and vision, were screened for color blindness (Ishihara Colorblindness Test; Clark, 1924), and had full motor control of both hands and at least one foot. All participants self-reported they learned English prior to the age of one year-old and could read and speak English fluently.

2.2 Hardware and Software

Visual stimuli were presented using an ASUS Intel Core i5 desktop computer. Participants responded to experiment-based events using a Cedrus RB-840 response box and/or an iKKEGOL USB Single Foot Switch Control One Key Customized Computer Keyboard Action Pedal HID. Participants pressed the spacebar on a standard keyboard to proceed after a short break between conditions. We used Python (Van Rossum and Drake, 2009) to build the experiment and R (RStudio Team, 2022) to analyze the data.

2.3 Tasks

We investigated the influence of visual load (number of targets, distractor [i.e., friendlies] present/absent) and overall demands (one or two tasks performed simultaneously) on participants' efficiency to respond to critical events in a multiple object tracking (MOT) task and/or a detection of electronic attacks (EA) task. In both tasks, participants were asked to monitor the targets displayed on a single visual display. Participants could identify targets using shape (circle), color (red), size (larger than friendlies), and/or speed (faster than friendlies). The shapes (octagon, diamond) and colors (pink, magenta) of friendlies were chosen using the psychophysical results of previous research of Glavan and colleagues (2020). The speed and size of tracks were chosen using pilot tests and practical considerations; size: displayed non-overlapping tracks on screen simultaneously (up to 1000 tracks), speed: balanced participants ability to track multiple moving targets (slower is more manageable) while providing an adequate number of instances that demanded a response from the participant (faster leads to more observations). The attributes of this graphical radar interface are not equivalent to the ACS but serve as a laboratory instantiation of the stressors placed on cognitive processing of MOT and detection tasks in Aegis-like contexts. Example stimuli are provided in Figure 1.

The cognitive and response demand were different for the MOT and EA tasks (described in Section 2.3.1 and 2.3.2, respectively). However, the participants' perceptual demands were consistent across tasks – stimuli were embedded within the same visual display, and both required participants to track and make decisions about hostile tracks while ignoring friendlies. For both tasks, perceptual demands systematically varied across conditions, i.e., the number of simultaneously presented targets (2, 4, or 6 hostiles) and the presence or absence of friendlies (see Figure 2 for examples).

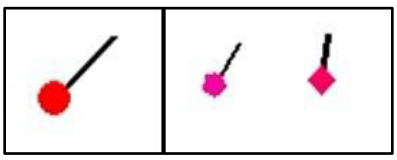


Figure 1. Example stimuli, enlarged. A hostile (target) track (left) and two types of friendly (distractor) tracks (right).

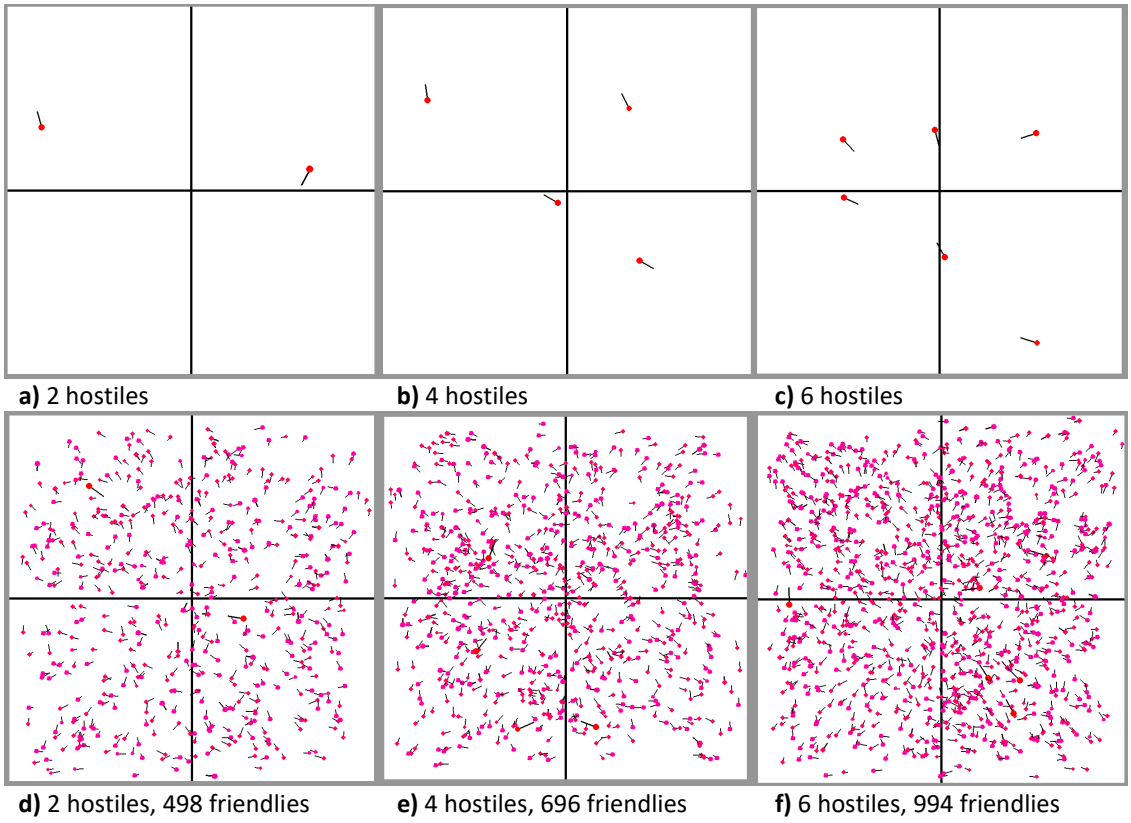


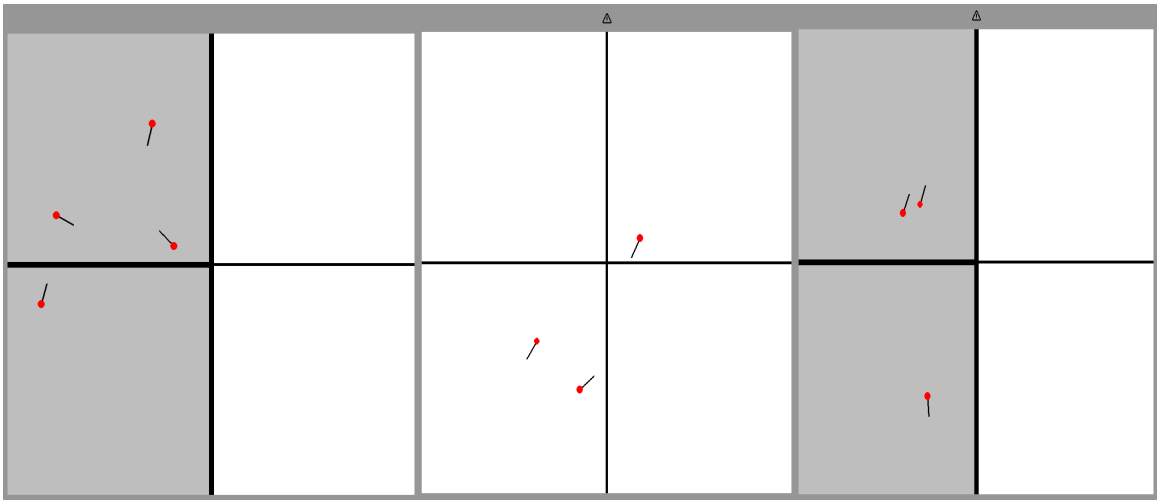
Figure 2. Example displays by factorial combination of number of targets (2, 4, or 6) and absence (subfigure 2a-2c, respectively) or presence (subfigure 2d-2f, respectively) of friendlies.

2.3.1 Multiple object tracking (MOT)

Participants were instructed to closely track hostiles (large red, fast-moving, circles) and ignore friendlies (smaller pink/magenta, slow-moving, octagons/diamonds). Once located, participants used 4 buttons on a response pad to mark quadrants with one or more hostile targets within its bounds, referred to as 'alarm ON'; each button was mapped to either the northeast (right middle finger), southeast (right index finger), northwest (left middle finger), or southwest (left index finger) quadrant. Quadrants with an alarm ON had a gray, as opposed to white, background (Figure 3a). Participants were also responsible to turn OFF the alarm when all hostiles left the quadrant; participants turned OFF an alarm by pressing the corresponding button on the response pad. Participants' accuracy and time to respond to hostiles entering (alarm ON) and leaving (alarm OFF) quadrants was recorded.

2.3.1 Electronic attacks (EA)

Identical to the MOT task, participants were instructed to closely track hostiles (large red, fast-moving, circles) and ignore friendlies (smaller pink/magenta, slow-moving, octagons/diamonds). Once located, participants tracked the hostiles and indicated when one disappeared. An EA occurred for a brief (1-4 seconds) time, once every 30 seconds and at a random time within each non-overlapping interval. The EA caused a single hostile track to disappear from the screen without warning. The participant was instructed to respond by pressing a foot pedal as quickly as accurately as possible when they detected an EA started (i.e., a hostile disappeared) and ended (i.e., a hostile reappeared). The affected hostile track would continue moving and reappear in its new location. The EA task did *not* depend on the quadrant that the hostile was in, rather it was a single alert that was turned ON and OFF by a single foot pedal. As such, an active EA alert was indicated by a visual icon located to the North of the quadrants (Figure 3b-c). Participants' accuracy and time to respond to EA (start, end) was recorded.



a) MOT task, alarms ON **b)** EA task, alarm ON **c)** MOT and EA task, alarms ON

Figure 3. Example display of the appropriate alarms turned ON in the MOT task (3a), EA task (3b), or both (3c).

2.4 Procedure

This within-subjects experiment was 5.5 hours, across three sessions. The first 30 minutes of Session 1 included COVID-19 requirements (i.e., Rapid Response testing, self-reporting vaccination status), the review of and consent to participate, the completion of a demographic survey, and a test for color blindness. The remaining time of Session 1, and all following Sessions, took 1.5 hours to complete. In Session 1, participants were introduced to the MOT task, and told they were to act as a radar operator. They were responsible for alerting their team where hostile targets were located (described in Section 2.3.1.). After adequate training and practice in all conditions – with(out) distractors, and 2, 4, or 6 hostiles – participants performed 6, 13-minute MOT task conditions (Figure 2). All conditions *without* distractors (Figure 2a-2c) were completed first, in a randomized order, followed by conditions with friendlies (Figure 2d-2f), in a randomized order. Participants saw their rough performance score (points based on time and accuracy to respond) at the end of each condition. No incentives were given to participants regarding their score. Participants took a 15+ second mandatory break between conditions (pressing the spacebar to continue). After all conditions, participants were asked to describe their strategy (if any) to complete the task (free response). Identical procedures were repeated in Session 2, however, participants were introduced and trained to complete the EA task (described in Section 2.3.2). In Session 3, procedures were repeated, but participants completed both tasks simultaneously. Participants were compensated for their time after each session and given the completion bonus with their payment for Session 3. The experiment concluded with a debrief of our motivation for research and we thanked them for their participation.

2.5 Analyses

We used a nonparametric measure of performance to quantify how well participants performed each task independently, referred to as the *Cost Coefficient* (*c*), and both tasks together, *MT*. These measures provided us a quantitative and qualitative interpretation of efficiency, and provided a standardized and individualized, model-based coefficient of the full function of efficiency to appropriately capture variation across multiple independent variables. We provided a summary of the measures here; an interested reader can refer to Fox et al. (2021) for a full derivation and demonstration of the measures. *Cost* was calculated using a transformation of the distribution of response times, called the cumulative reverse hazard function, $K(t)$, which is the cumulative of the reverse hazard function, $k(t)$. The reverse hazard function is defined as the ratio of the probability density function (PDF) and the cumulative distribution function (CDF), $k(t) = f(t)/F(t)$. The *Cost* (Equation 1) of a manipulation on each task was calculated across a full range of performance, rather than just the mean (a common practice in psychological research), i.e., t_{min} - t_{max}.

$$c_i = \int_{t_{min}}^{t_{max}} k_i(t) - k_{i(N)}(t)\, dt = K_i(t_{max}) - K_{i(N)}(t_{max})$$

(Equation 1)

Then, we calculated MT using a weighted sum of the dual-task *Cost* for each task, where the sum of the weights equals 1, $\sum^{N}_{i=1} w_i = 1$:

$$\mathrm{MT} = \sum_{i=1}^{N} w_i c_i. \qquad \text{(Equation 2)}$$

In this application, we assumed equal weight to the MOT and EA tasks, following the instruction provided to participants. Both *Cost* and *MT* provide a function of efficiency across all times, *t*, so we calculated their functional *z*-score to serve as a summary coefficient; this allowed for a straightforward way to compare efficiency across our independent measures of interest. For instance, we may find a negative MT coefficient for all dual-task conditions, evidence of a limited function of MT (as expected), but perhaps some conditions are a magnitude more/less efficient (double or triple) than others. The z-score coefficient is interpreted as the number of standard deviations (SD) that observed performance under our independent variable(s), $K_{i(n)}(t)$, falls below or above the baseline performance, $K_i(t)$, and it is representative for the full function of times, *t*. If the *z*-score is above zero, efficiency is labeled as *super capacity*; if it is equal to zero it is labeled as *unlimited capacity*; if it is below zero it is labeled as *limited capacity*. This powerful analytic modeling approach allowed us to make comparisons across the same scale of measurement, and at the functional- and individualized-level.

3. Results

First, we report the summary statistics of task performance by condition, at the group-level (Table 1). Overall, we found expected patterns of variation in performance. Specifically, performance (response time [RT], accuracy) decreased as the number of targets increased and in the presence (versus absence) of friendlies, across both the MOT and EA tasks. Additional performance deficits were observed in both tasks when performed together (dual-tasking) versus each in isolation (single-tasking).

Table 1: Effect of independent variables on mean (SD) of individuals' performance by task.

	MOT		EA	
	RT (SD)	Accuracy (SD)	RT (SD)	Accuracy (SD)
2	796 (1035)	92.3 (26)	851 (739)	96.5 (18)
4	976 (1198)	85.5 (35)	914 (924)	89.5 (31)
6	1021 (1276)	79.3 (41)	917 (1027)	78.3 (41)
w/o Friendlies	723 (828)	88.0 (33)	840 (792)	96.3 (19)
w/ Friendlies	1144 (1420)	83.2 (37)	962 (1004)	79.7 (40)
Single-task	891 (1115)	86.6 (34)	702 (653)	88.0 (33)
Dual-task	971 (1038)	84.6 (36)	1080 (1038)	89.5 (31)

Note. The mean (SD) of participants' response time (RT; milliseconds) and accuracy (percent correct) by task (MOT, EA) and experimental condition: # of hostiles, presence (absence) of friendlies, and single- or dual-task. SD = standard deviation; MOT = multiple object tracking; EA = electronic attack; # = number; w/o = without; w/ = with.

Next, we report the summary statistics of individuals' Cost in efficiency (Table 2) and MT coefficients (Table 3) by condition. Figure 4 illustrates Cost (or lack thereof) of task performance due to tracking more than 2 hostiles, the presence of friendlies, and/or dual-tasking, calculated using Equation 1, for each task. Individuals' *z*-score Cost coefficients are summarized (mean, standard deviation) in Table 2. These Cost coefficients describe the degree of difference between the cumulative reverse hazard functions for each independent variable, standardized by an individuals' reverse hazard function of performance on each respective task in isolation while tracking 2 hostiles without the presence of friendlies, i.e., customized individual-based baselines.

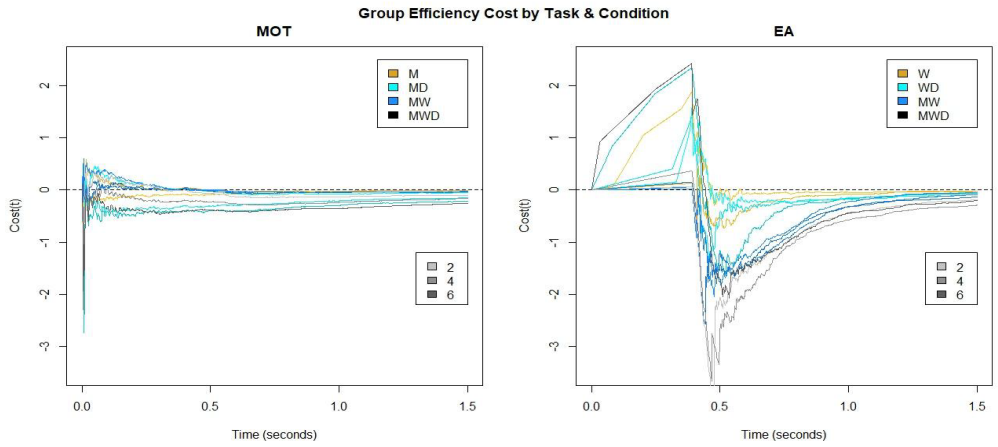


Figure 4: Group Cost as a function of time to respond by task (MOT, EA) and experimental condition: # of hostiles, single- or dual-task, and presence (absence) of friendlies. MOT = multiple object tracking; EA = electronic attack; M = MOT alone, W = EA alone; D = distractors; WD = EA with friendlies; MD = MOT with friendlies, MW = dual-task; MWD = dual-task with friendlies; 2, 4, 6 = number of hostiles to track; # = number.

Table 2: Effect of independent variables on mean (SD) of individuals' Cost coefficients (*z* score) by task.

		Single-task		Dual-task	
Task	**# of Hostiles**	**w/o Friendlies**	**w/ Friendlies**	**w/o Friendlies**	**w/ Friendlies**
MOT	**2**	Baseline	-0.08 (1.36)	0.26 (1.67)	-0.24 (1.95)
	4	-0.92 (1.29)	-2.17 (1.38)	-0.64 (1.91)	-1.66 (1.62)
	6	-0.43 (1.53)	-2.53 (1.28)	-0.45 (1.90)	-2.22 (1.78)
EA	**2**	Baseline	-0.64 (1.23)	-2.88 (0.76)	-3.25 (0.78)
	4	-0.44 (1.60)	-0.94 (1.32)	-2.98 (0.91)	-2.95 (0.83)
	6	-1.25 (1.34)	-1.47 (0.83)	-2.84 (0.89)	-1.88 (1.45)

Note. The mean (SD) in individuals' Cost by task (MOT, EA) and experimental condition: # of hostiles, single- or dual-task, and presence (absence) of friendlies. SD = standard deviation; MOT = multiple object tracking; EA = electronic attack; # = number; w/o = without; w/ = with.

We combine the Cost for each of the tasks to obtain a MT coefficient of the factorial combination of our dual-task conditions: with and without the presence of friendlies, and 2, 4, and 6 hostiles to track. Individuals' MT coefficients are summarized (mean, standard deviation) in Table 3.

Table 3: Effect of the number of hostiles to track and presence (absence) of friendlies on mean (SD) of individuals' MT coefficients.

# of Hostiles	**w/o Friendlies**	**w/ Friendlies**
2	-1.31 (0.89)	-1.74 (1.15)
4	-1.81 (1.13)	-2.30 (1.00)
6	-1.64 (1.10)	-2.05 (1.18)

Note. The mean (SD) in individuals' MT by experimental condition: # of hostiles, and presence (absence) of friendlies. MT = Multitasking Throughput; SD = standard deviation; # = number; w/o = without; w/ = with.

4. Discussion

Our goal was to quantify the variation in human efficiency to perform Aegis relevant tasks under contextual and cognitive stressors which influence attention and performance. We computed coefficients for each individual and condition. We report (illustrate) a summary of our results using functional *z*-score coefficients. In a multisession, controlled laboratory experiment, we asked people to complete a MOT task and/or detect instances of one type of EA on hostile entities, known as jamming. In both tasks, we asked people to monitor the location of hostiles (red circles) in isolation or embedded within several hundred friendlies (i.e., civilians). Friendlies were smaller in size and traveled at slightly slower speeds than hostiles but resembled threats in color (pink, magenta) and shape (octagons, diamonds). The heading direction of an entity could change at any given time, and, in instances of EA, a hostile entity could disappear for a few moments. These factors increased uncertainty in the projected travel path of entities and required participants to continuously track and monitor each hostile (target) entity. We asked participants to track hostiles and ignore friendlies while responding to critical events. Participants pressed buttons with their hands (MOT task) or feet (EA task) to respond when hostiles changed critical locations (i.e., entered/exited a quadrant) or disappeared/reappeared from the radar, respectively.

4.1 Analysis of Variance, Cost, and MT

As expected, we found that overall performance decreased with a larger number of hostiles to track, the presence of friendlies, and when given the responsibility to both track entity location and detect instances of EA – that is, response times increased and accuracy decreased. The model-based and individualized measures we applied, Cost and MT, provided a more rigorous investigation of the effect of our independent variables on individuals' performance, compared to their baseline performance, and across the full function of response times, rather than just the mean. We defined individuals' baselines using the cumulative reverse hazard function of their time to make responses when only two hostiles were to be tracked; an independent baseline was calculated for the MOT and EA task, and each was performed in an isolated condition.

Giving equal weight to the MOT and EA tasks, we computed individuals' MT under each tracking load and with(out) the presence of friendlies. We found individuals' MT was limited in all dual-task conditions, and MT suffered slightly more with the presence of friendlies and when tracking more than two hostiles. More specifically, we found peoples' efficiency, calculated using the Cost metric, to turn ON/OFF alarms (MOT task) decreased between 1.6 and 2.5 SDs in the presence versus absence of friendlies (i.e., limited capacity), but only when the number of hostiles to track increased above two. Participants' efficiency to perform the MOT task was not affected by the presence of the EA task or in any condition where they were responsible to track only two hostiles (i.e., unlimited capacity). In the absence of friendlies, there was only a marginal effect of the number of hostiles on efficiency.

Compared to MOT events, we found opposite effects of our manipulations on peoples' efficiency to detect the (dis)appearance of a hostile, as calculated using the Cost metric to EA task events. Specifically, we found efficiency in the EA task suffered most dramatically in the presence of the MOT task, and this was the primary driver of the decrease observed in participants' efficiency in the EA task, regardless of the presence (absence) of friendlies or the total number of hostiles to track. In other words, the presence of the MOT task evoked very limited capacity for individuals to detect and respond to EA events in *all* conditions. Further, in the absence of the MOT task (i.e., EA alone) the presence of friendlies did not have any practical effect on EA efficiency alone (unlimited capacity), but the number of targets did have some degree of effect on efficiency and did so in an expected way.

4.2 Implications, Limitations, and Future Research

An ACS can be intelligently designed to assist operators to efficiently track, monitor, and make decisions about threats. Our work demonstrates that the degree of deficit (if any) each cognitive manipulation has on efficiency depends on the nature of the task(s). We utilized two tasks that required continuous tracking of hostiles, but they differed in their underlying cognitive demands. We found this difference evoked unique patterns of efficiency. Specifically, the efficiency of participants to report hostiles' locations suffered when simultaneously tracking more than two, and efficiency decreased most drastically when hostiles were embedded within hundreds of closely resembling friendlies. When this type of task is crucial for mission success, designers may prioritize their focus on providing a large visual discrepancy between hostiles and friendlies rather than decreasing the total number of hostiles to track or to offload dual-task responsibilities. This is in line with

previous research – peripheral vision and quick eye movements are more susceptible to visual crowding, especially for perceptually similar and moving entities (Kooi et al., 1994). Hence, the largest efficiency deficit was observed in the MOT task when friendlies were present, and we expect this finding to generalize to similar tasks that rely on peripheral and unfocused vision, and, when tracking more than two moving entities.

On the other hand, the presence of the MOT task had the largest influence on EA efficiency. Here, we suspect a strategic shift in how individuals decided to track hostile entities when performing the EA task alone versus with the MOT task. The detection of hostile (dis)appearance required foveal tracking of each hostile, but in dual-task conditions participants reported that they attempted to group hostiles to attain both their location (MOT task) and presence (EA task). This necessarily spread their focal visual attention among several entities, and groups of entities, and led to decreased processing efficiency of each individual entity (i.e., limited capacity processing), a pattern of performance that aligns with limited resource theories of attention (Wickens, 1984; 2002). Further, our participants reported focusing on the general AoI where hostiles were located (MOT) rather than counting hostiles and detecting their (dis)appearance (EA). More generally, we suggest a task requiring foveal vision may suffer when paired with that which demands peripheral vision, as it is unintentionally underprioritized. Future work should investigate whether performance deficits may be partially alleviated when two tasks that require foveal vision (and are embedded within the same framework) are paired; a somewhat counterintuitive prediction based on theories of competition for shared cognitive resources. For example, future research could assess individuals' efficiency when attempting to detect both the disappearance of existing targets and the addition of new targets.

5. Conclusions

In this paper, we modeled the influence of external factors on cognitive load and performance in radar tasks. We captured how, and the degree to which, variation in efficiency changed as participants attempted to track and make decisions about the location and presence of hostiles on a radar display. We found the number of targets to track and the presence of friendlies highly affected efficiency in a MOT task, one that primarily required peripheral vision. Alternatively, we found that the presence of the MOT task substantially influenced efficiency to detect instances of EA, and this was persistent across all conditions. The analytic method we applied accounted for the full function of RTs and allowed us to identify the relationship between meaningful factors that impact performance in operational environments. Our discussion gives fodder to system designers seeking to research and develop tools that may alleviate cognitive demands and/or counter threats of EA in radar tasks. We closed with a few potential directions for future research.

Acknowledgements

The views expressed are those of the authors and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government. No potential conflict of interest was reported by the authors. The study was approved by the AFRL 711 HPW IRB (FWR20220072E). Distribution A. Approved for public release; distribution unlimited. AFRL-2022-5531; Cleared 16 Nov 2022.

References

Altmann, E. M. and Trafton, J. (2004) Task Interruption: Resumption Lag and the Role of Cues. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Lansing, Michigan, August.

Bath, W. G. (2020) Overview of Platforms and Combat Systems. *Johns Hopkins APL Technical Digest,* Vol 35, No. 2. Available from: www.jhuapl.edu/Content/techdigest.

Blair, W. D., Watson, G. A., Kirubarajan, T. and Bar-Shalom, Y. (1998) Benchmark for radar allocation and tracking in ECM. *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 34, No. 4, pp. 1097-1114. Available from: doi.org/10.1109/7.722694.

Clark, J. H. (1924) The Ishihara Test for Color Blindness. *American Journal of Physiological Optics,* Vol. 5, pp. 269-276.

Fisher, C.W. and Kingma, B. R. (2001) Criticality of data quality as exemplified in two disasters. *Information and Management,* Vol. 39, pp. 109-116. Available from: doi.org/10.1016/S0378-7206(01)00083-0.

Fox, E. L., Houpt, J. W. and Tsang, P. S. (2021) Derivation and demonstration of a new metric for multitasking performance. *Human Factors*, Vol. 63, No. 5, pp. 833-853. Available from: doi.org/10.1177/0018720820951089.

Glavan, J.J., Haggit, J.M. and Houpt, J.W. (2020) Temporal organization of color and shape processing during visual search. *Attention, Perception, and Psychophysics,* Vol. 82, pp. 426-456. Available from: doi.org/10.3758/s13414-019-01863-9.

Gurbuz, S. Z., Griffiths, H. D., Charlish, A., Rangaswamy, M., Greco, M. S. and Bell, K. (2019) An overview of cognitive radar: Past, present, and future. *IEEE Aerospace and Electronic Systems Magazine,* Vol. 34, No. 12, pp. 6-18. Available from: dio.org/10.1109/MAES.2019.2953762.

Joint Chiefs of Staff. 2013. *Electronic Warfare* (JP 3-13.1). [Online]. Government report. Available from: info.publicintelligence.net/JCS-EW.

Kang, H. (2021) Sample size determination and power analysis using the G* Power software. *Journal of educational evaluation for health professions,* Vol. 18. Available from: doi.org/10.3352/jeehp.2021.18.17.

Kooi, F. L., Toet, A., Tripathy, S. P. and Levi, D. M. (1994) The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial vision,* Vol. 8, No. 2, pp. 255-280. Available from: doi.org/10.1163/156856894X00350.

Morgan, B., D'Mello, S., Abbott, R., Radvansky, G., Haass, M. and Tamplin, A. (2013) Individual differences in multitasking ability and adaptability. *Human factors,* Vol. 55, No. 4, pp. 776-788. Available from: doi.org/10.1177/001872081247084.

Ritter, F. E., Tehranchi, F. and Oury, J. D. (2019) ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science,* Vol. 10, No. 3. Available from: doi.org/10.1002/wcs.1488.

Park, S., Nam, I. and Noh, S. (2018) Modeling and Simulation for the Investigation of Radar Responses to Electronic Attacks in Electronic Warfare Environments. *Security and Communication Networks*, pp. 1-13. Available from: doi.org/10.1155/2018/3580536.

Pichlmeier S. and Pfeiffer T. (2021) Adaptive target enhancement determines levels of guidability in Multiple Object Tracking. *Vision Research,* Vol. 183, pp. 61-72. Available from: doi.org/10.1016/j.visres.2021.02.001.

Porgue, D. (2016) "5 of the Worst User-Interface Disasters". *Scientific American.* April.

RStudio Team. (2022) RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA.

Spezio, A. E. (2002) Electronic warfare systems. *IEEE Transactions on Microwave Theory and Techniques*, Vol. 50, No. 3, pp. 633-644. Available from: doi.org/10.1109/22.989948.

Van Rossum, G. and Drake, F.L. (2009) Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

Walsh, M. M., Gunzelmann, G. and Dongen, H. P. (2017) Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review*, Vol. 24, No. 6, pp. 1785-1807. Available from: doi.org/10.3758/s13423-017-1243-6.

Nuclear Cyber Attacks: A Study of Sabotage and Regulation of Critical Infrastructure

Virginia A. Greiman
Boston University, MA, USA

ggreiman@bu.edu

Abstract: As of 2021, the World Nuclear Association reports 440 Nuclear reactors are in operation worldwide in 30 countries generating capacity of 390 (GW) which is equivalent to about 10% of the world's electricity. After Hydroelectric power, nuclear is the world's second largest source of low-carbon power. Important new nuclear technologies including the Small Modular Reactors (SMRs) are being developed globally creating more efficient and safer reactors that can be reproduced off site. While governments redoubled their commitments to reducing greenhouse gas emissions at the UN Climate Change Conference (COP26) in Glasgow, the recovery of economies following the harsh impacts of COVID-19 led to a surge in energy demand that surpassed the growth in production from clean sources including nuclear. The safety and security of nuclear power has received renewed attention since the Russian invasion of Ukraine presenting growing concern about the potential threat of increased malevolent cyber activity against Ukraine's critical infrastructure. Moreover, there have been more than 20 known cyber incidents worldwide at nuclear facilities since 1990. To address these concerns this paper focuses on the progress of cyber security and cyber resilience in the nuclear industry globally. The 2015 cyber-attack on the Ukrainian Kyivoblenergo, a regional electricity distribution company was analyzed by multiple sources including private companies, investigators in Ukraine, and the U.S. government. The analysis revealed many opportunities to stop or prevent this attack, however, the nuclear industry continues to face serious challenges in protecting against cyber threats. This research will investigate through a comparative analysis the recent government regulations, rules and standards, for nuclear cyber security safety in the United States and internationally to determine whether these laws adequately protect energy infrastructure from cyberattacks and hold responsible parties accountable. Recent initiatives by government and the private sector to enhance the opportunities for improving cyber security in the nuclear sector will be reviewed to determine best practices for improving nuclear safety and cyber resilience.

Keywords: Nuclear Cyber Security, Cyber Resilience, Nuclear Regulation, Nuclear Power Plant Cyberattacks

1. Introduction

The recent cyber-attacks on a nuclear facility in the United States and Saudi Arabia, highlights the vulnerabilities of nuclear facilities. On March 24, 2022, the Department of Justice released sealed indictments against four Russian officials who hacked into the Wolf Creek operating system, Kansas' first nuclear power generating station and compromised a petrochemical facility in Saudi Arabia. According to the Indictment, the group used their hacking skills in two separate conspiracies which targeted the global energy sector between 2012 and 2018 (DOJ 2022). "Russian state-sponsored hackers pose a serious and persistent threat to critical infrastructure both in the United States and around the world," said Deputy Attorney General Lisa O. Monaco. "Although the criminal charges reflect past activity, they make crystal clear the urgent ongoing need for American businesses to harden their defenses and remain vigilant" (DOJ 2022).

According to the indictments, the four officials, including three members of Russia's domestic intelligence agency, the Federal Security Service, or F.S.B., are accused of breaching hundreds of energy companies around the world, showing the "dark art of the possible," a Justice Department official said at a briefing with reporters. Investigators believed at the time that the intrusion was meant to trigger an explosion, but said that a mistake in the code prevented one (DOJ 2022). Undeterred, the next year the hackers researched refineries in the United States and tried to breach the computers of an American company that managed similar critical infrastructure facilities in the United States, according to court filings. They used several tactics to gain access to computer networks, including spear phishing attacks that targeted more than 3,300 users at more than 500 American and international companies (DOJ 2022). They targeted government agencies such as the Nuclear Regulatory Commission, and in some cases they were successful.

The cyber-attack in India's largest civil nuclear facility – the Kudankulam Nuclear Power Plant (KNPP) in Tamil Nadu in September 2019 highlights the seriousness of cyber-attacks, and raises new concerns about the vulnerability of nuclear plants to serious destruction and potential loss of life (Chaudhury, 2019). Malware was discovered in systems used to manage administrative activities at the Kudankulam nuclear power plant and the attack was linked by the Nuclear Power Corporation of India to the Lazarus Group, operating from North Korea (Campbell and Singh 2019). Though the KNPP attack was not intended to cause destruction but to extort the confidential data and reconnaissance, future attacks may have more serious consequences.

Cyber-attacks provide a new opportunity for hostile powers to cause mayhem at nuclear facilities without ever having to have direct access to the facility. The compromise of networks in nuclear facilities can be used to facilitate sabotage, malicious acts involving nuclear material, compromise sensitive data, and create a reactor catastrophe. This is of grave concern in densely populated countries like India, as any radiation release from a nuclear facility would cause tremendous damage and loss.

The contributions of this paper are two-fold. First, this research discusses the increasing risk that nuclear plants face and the urgent need for a focus on stronger oversight of the nuclear industry that places responsibility on the parties best positioned to control the risk of a cyber-attack. Second, to review the cyber security regulatory environment specific to the nuclear industry followed by a discussion and recommendations on the initiatives that governments and the private sector – can and should take - in reducing the likelihood and impact of cyber-attacks on nuclear facilities.

2. Cyber security risk and nuclear power

The recent cyber-attacks in the nuclear industry reflect a growing threat to nuclear facilities globally. According to the U.S. Department of Energy, the past 20 years have seen the threat from malicious cyber attackers increase substantially, powered by new incentives for conducting attacks, such as drawing a ransom payment (DOE 2022). Cyber-attacks could be associated with business espionage, technology theft, a disgruntled employee, a recreational hacker, a cyber activist, organized crime, a nation state, or a terrorist organization (Arinze, et al. 2020). The overall expansion of the cybersecurity threat is a rising tide for all sectors; however, it is especially important to the energy sector (ODNI 2021). Lloyd's City Risk Index 2015-2025 reports due to an increasingly interconnected and technologically dependent world, that nearly half of the total GDP at risk is linked to manmade threats, including market crash, human pandemic, cyber-attack, power outage and nuclear accident (Lloyd's 2015).

The United States (U.S.), Russia, United Kingdom (UK), South Korea, and China have raised concerns about securing their facilities from catastrophic incidents. The Chernobyl incident of 1986 (WNA 2022a) and the Fukushima Daiichi nuclear disaster of 2011 (WNA 2022b) have also shown that disastrous consequences that can occur when the proper safety and security protocols are not followed (Kobayashi 2019). These events are indications of weak security and safety controls, resulting in reputation damage, loss of trust, reduction in shareholder value, financial fallout, and loss of human lives.

The most important security responses must be national and unilateral, focused on hygiene, redundancy, and resilience (Nye 2023). It is likely, however, that major governments will gradually discover that cooperation against the insecurity created by nonstate actors will require greater priority and attention. The world is a long distance from such a response at this stage in the development of cyber technology (Nye 2011).

3. Global sources of nuclear power

After Hydroelectric power, nuclear is the world's second largest source of low-carbon power. According to the World Nuclear Association (2023a) 13 countries in 2020 produced at least one-quarter of their electricity from nuclear. France gets around 70.6% of its electricity from nuclear energy, Slovakia and Ukraine get more than half from nuclear, while Hungary, Belgium, Slovenia, Bulgaria, Finland, and Czech Republic get one-third or more (WNA 2023a). South Korea normally gets more than 30% of its electricity from nuclear, while in the US, UK, Spain, Romania, and Russia about one-fifth or 20% of electricity is from nuclear.

Beyond power generation, nuclear technologies have medical applications that will help combat serious viruses like COVID 19. The International Atomic Energy Agency (IAEA) is providing diagnostic kits, equipment, and training in nuclear-derived detection techniques to countries asking for assistance in tackling the worldwide spread of the novel coronavirus causing Covid-19.

The International Thermonuclear Experimental Reactor (ITER) is presently building the world's largest and most advanced experimental tokamak nuclear fusion reactor in Provence, southern France (ITER 2022). ITER is an international nuclear fusion research and engineering megaproject. This is a true international endeavor run by seven member entities: the European Union, China, India, Japan, Russia, South Korea, and the United States. Overall, 35 countries are participating in the project directly or indirectly. The project was initiated in 1988 and is expected to start full deuterium-tritium fusion experiments in 2035.

Recently, as an alternative to large reactors, the U.S., the U.K. and Canada, three major nuclear markets, have all signaled growing support for the small modular reactors (SMRs) currently in development. Canada, for example, launched a 27-point SMR national action plan to demo and deploy the technology, update regulations and create employment. The global market for SMRs is expected to be worth up to $300 billion a year by 2040 (WEF 2021).

4. Nuclear cyber security regulations

Even though there has been some progress in developing enhanced cybersecurity measures across the nuclear industry, the research highlights the fragmentary and inconsistent response to cyberthreats at nuclear facilities by national governments and private-industry (Brunt and Unal 2019).

The 2020 Nuclear Security Index, published by the Nuclear Threat Initiative (NTI), shows progress on global nuclear security has slowed significantly over the past two years, despite sizeable gaps that continue to leave nuclear materials and facilities vulnerable to theft and acts of sabotage (NTI 2020).

According to the Index, regulatory requirements for nuclear security are not comprehensive, with significant weaknesses in important areas such as insider threat prevention, security culture, and cybersecurity, leaving dangerous gaps and vulnerabilities around the world. The Nuclear Security Index (NTI) found that 25% of countries with nuclear reactors don't have basic cybersecurity measures in place demonstrating an inconsistent global approach to cybersecurity programs at nuclear facilities (NTI 2020). Based on a country by country assessment NTI strongly recommended that significant steps be taken to improve cyber security at nuclear facilities thus reducing the potential for serious nuclear catastrophic events (NTI 2020). These include putting in place regulatory structures to implement national security, increasing global participation by strengthening global security, sharing information, and developing adequate regulatory and human resources to manage client countries' nuclear security responsibilities (NTI 2020).

4.1 United States nuclear security requirements

A year after the 9/11 attacks, the Nuclear Regulatory Commission (NRC) issued an order that included cyberattacks among the threats that nuclear plants would be required to defend against. Additional guidance for dealing with cyber threats was released during the next several years, and NRC issued formal cybersecurity regulations in March 2009 ("Protection of Digital Computer and Communications Systems and Networks," 10 CFR 73.54). NRC published a regulatory guide for the program in January 2010 (NRC 2010) (and in 2022 a proposed revision has been noticed) (NRC, 2022). Later that year, the Nuclear Energy Institute (NEI) also published implementing guidance in NEI 08-09, "Cyber Security Plan for Nuclear Power Plants" (and in 2018, an addendum). These documents provide information to aid licensees in developing cybersecurity plans. NRC's cybersecurity regulations require each nuclear power plant to submit a cybersecurity plan and implementation schedule. The plan must provide "high assurance" that digital computer and communications systems perform certain safety and security functions to provide adequate protection against design basis attacks.

NRC began inspecting the implementation of nuclear plant cybersecurity plans in January 2013. The inspections are part of NRC's Cyber Security Oversight Program, which is being incorporated into the existing Reactor Oversight Program (NRC 2020).

The Energy Policy Act of 2005 as amended (EPACT05, P.L. 109-58) imposed specific criteria for the NRC to consider in revising the "Design Basis Threat" (DBT), which specifies the maximum severity of potential attacks that a nuclear plant's security force must be capable of repelling. Among other changes, the revisions expanded the assumed capabilities of adversaries to operate as one or more teams and attack from multiple entry points (Holt and Andrews 2014)

Nuclear power plants are protected from cyberattacks by layers of safety precautions, with the first line of defense being isolation (NEI 2016). This means they have no direct access to web, nor do they have indirect access because they are not connected to the plants' internal networks" (NEI 2016). Some steps nuclear power plants have taken to protect against cyber threats include: "[i]solat[ing] key control systems;" "[e]nhanc[ing] and implement[ing] strict controls over the use of portable media and equipment;" and "[p]erform[ing] detailed cyber security assessments." (NEI 2016). Despite all of the layers of protection in the nuclear industry, it is still vulnerable to cyberattacks.

Notably, existing U.S. nuclear power reactors, designed in the 1960s and '70s, are controlled primarily by analog systems that are resistant to cyberattack. However, new reactors are being designed with digital controls, and

existing analog plants increasingly rely on digital computers to run auxiliary and monitoring systems. This increasing use of digital systems in nuclear power plants, along with post 9/11 security concerns and at least one "worm" infection at a U.S. reactor, (Kesler 2001) have prompted increased NRC attention to cybersecurity.

4.2 Global cyber regulation of nuclear power

Beyond the United States, the topic of cybersecurity in nuclear programs has been examined at the international scale as well (Pickering and Davies 2021). The Nuclear Threat Initiative (NTI) "cyber score index" illustrates the wide range of how seriously cyber threats are being addressed, from countries with well-developed nuclear power capabilities, such as the United States, Canada, France, Switzerland, and Russia, to countries with nuclear reactors and much more limited oversight, such as Mexico, Brazil, Italy, Kazakhstan, and China (NTI 2020). The NTI identified four overarching priorities they believe would substantially reduce the risk of damaging cyber-attacks on nuclear facilities. The priorities include institutionalizing cyber security, mounting an active defense, reducing complexity, and pursuing transformation. These recommendations are ambitious and not easily achieved, but if implemented, they would dramatically reduce the probability of a successful cyber-attack (NTI 2020).

One of the common threads among countries with limited nuclear capabilities according to NTI's 2020 Cyber Score Index is the lack of an institutional body or organizational structure focused on cyber security threats in the nuclear industry. For example, Ankara, Turkey is still trying to prepare necessary regulations and legislations to be ready for a proper establishment of the nuclear facility's infrastructure (NTI 2020). All ministries and public offices are approaching the problem from a micro perspective and are regulating their areas of interests. However, there is no coordinating authority to concentrate these micro perspectives into a macro one (NTI, 2020). Secondly, Turkey has no Industrial Control System (ICS)-specific cyber security organization, which could coordinate the private and state stakeholders in the sector (NTI 2020). By considering the recent political developments in Turkey and the ambiguity of international law on cyber-attacks, Turkey has to develop its own defensive and offensive cyber security capacity. Ankara has to persistently focus on coordination and strategic communication among necessary parties (NTI 2020).

To address the problem faced by Turkey and other countries with limited capacity to manage large nuclear facilities, the International Atomic Energy Agency (IAEA) and other international organizations are actively working to provide guidance and recommendations for cyber protection of nuclear facilities (IAEA 2021). Nuclear security is the responsibility of each individual country, but international cooperation is vital to support States in establishing and maintaining effective nuclear security regimes.

The position of the International Atomic Energy Agency (IAEA) is that security is a national responsibility (IAEA, 2021). For the civil nuclear sector, the IAEA publishes nuclear security recommendations that reflect internationally accepted best practice, but there are no recognized international regulations as such. Other non-governmental organizations and lobbying groups, such as the World Institute for Nuclear Security (WINS, 2014), have issued high-level guidance documents on how national infrastructure for cybersecurity might be prepared, while instruction from organizations that set industry guidance such as the Nuclear Energy Institute (NEI) have also provided useful technical advice. Most recently, the Department of Homeland Security's Cybersecurity and Infrastructure Security Agency (CISA) has released numerous Alerts and Malware Analysis Reports regarding Russia's malign cyber activities, including the activities discussed in the indictments (CISA 2022). The National Institute of Standards and Technology (NIST), a non-regulatory agency of the US Department of Commerce, has developed guidance specifically to assist nuclear facilities in complying with multi-faceted regulations to verify that the computers, digital communications and systems in CNI are protected from cyberattacks (NIST 2018).

5. Recommended initiatives

An analysis of the laws, regulations and standards governing nuclear cyber security and safety reveals that better organizational structures and frameworks are needed to adequately protect energy infrastructure from cyberattacks. The failure of accountability for cyber-attacks is a gap that must be remedied through proactive and coordinated efforts between governments and the private sector that support more rigid standards, tighter security control systems, a national cyber oversight authority, and development and enforcement of cyber security policies and regulations.

Through this research recent initiatives and lessons learned by governments and the private sector to enhance the opportunities in the nuclear sector have been explored and various approaches considered. Based on this analysis, the following four initiatives are recommended for improving nuclear facilities cyber security strategy:

(1) A cyber resilient approach to recovery, (2) preparation for cyber-attacks through more engaged public private partnerships, (3) early detection through mindful organizing, and (4) understanding where security has failed and improving the security of nuclear cyber security systems.

5.1 Cyber resilience as a solution to recovery from nuclear cyber attacks

The looming threat of a cyber-attack on the nuclear industry that will cause serious harm is real and possibly imminent. A key concept that supports the survivability of a cyber-attack is resilience. Though often connected with cyber security, cyber resilience has a distinct and significant purpose. For too long, the focus of response to cyber-attacks across industries has been cyber security. This paper recommends that complementary to cyber security, cyber resilience be adopted as a critical approach to resolving cyber-attacks in the nuclear industry (Greiman and Bernardin 2021). Today's world complexity leads researchers and practitioners to focus on the ability of the organizations to recover from setbacks, adapt well to change, and keep going in the face of adversity. "It is not just the physical structure which must withstand a blow and come back, we need resilient staff, resilient management, resilient plans, and planning" (Manto and Lockmer 2015, p. 199).

The National Institute of Standards and Technology (NIST) defines cyber resilience as "the ability to anticipate, withstand, recover from, and adapt to adverse conditions, stresses, attacks, or compromises on systems that use or are enabled by cyber resources." The resilience of nuclear energy is the result of the combination of high levels of safety, operational flexibility and continuous learning from previous events. As recommended, by the Cyber Security and Infrastructure Security Agency (CISA), the organization's resilience to a destructive cyber incident must be maximized by test backup procedures to ensure that critical data can be rapidly restored if the organization is impacted by a destructive cyberattack (CISA, 2022).

As reflected in the OECD's Nuclear Energy Agency Policy (NEA, 2020) nuclear energy provides resilience on four main levels: (1) The first level is security by design. Design and beyond design nuclear power plants are conceived following the principles of defense-in-depth: prevention, protection and mitigation. This results in the implementation of redundant, independent and diversified safeguards designed to withstand external hazards. (2) The second level is the organizational level where resilience includes emergency preparedness, safety culture, continuous learning and international cooperation (Nucnet 2020). (3) The third level is the system level which has the following features: dispatchable and flexible; grid stability (inertia, reactive capacity, frequency control); on-site strategic fuel stockpiles; and simplicity of centralised systems. (4) The fourth Level is socio-economic, whereby there are competitive electricity prices; stable long-term investment; local high-paying jobs; and economic spill-over.

Resilience is a long-term endeavor that necessitates technology neutral policy actions beyond well-designed stimulus packages to properly guide investment decisions over time and correct potential market flaws (NEA 2020).

5.2 Preparation for cyber-attacks: Public private partnerships

In addition to resilient approaches to respond to cyber- attacks, a second important recommendation is to assess preparation for cyber-attacks at the national level, particularly through strong military offense and defense capabilities. According to the Defense Science Board (DSB 2017) as offensive cyber capabilities continue to grow, and are likely to outpace cyber defense, there are likely to be growing risks of misperception that could lead to rapid cyber escalation – and the potential for rapid escalation to armed conflict. In the United States, "[t]he U.S. Government must work with the private sector to intensify efforts to defend and boost the cyber resilience of U.S. critical infrastructure in order to avoid allowing extensive vulnerability to these nations.

The nuclear industry must also share information with other sectors. Without an ability to collaborate the nuclear industry cannot identify the attack patterns discovered by other sectors nor can it apply fool-proof security controls. It is no more palatable to allow the United States to be held hostage to catastrophic attack via cyber weapons by such actors than via nuclear weapons" (DSB 2017). Secondly, public-private partnerships for large-scale infrastructure projects have proven to be a formidable tool to reignite the economy and achieve social cohesion during previous crises (OECD 2019).

Early detection through mindful organizing

The wisdom of learning from failure is incontrovertible. Yet organizations that do it well are extraordinarily rare. High-reliability-organizations (HROs) are an example of organizations that have developed practices to help prevent catastrophic failures in complex systems like nuclear power plants through early detection. According to Weick & Sutcliffe, (2015) so-called high-reliability organizations (HROs) demonstrate particular characteristics in the way they operate: anticipating problems (being aware of what is happening in the work system; being alert to ways in which an incident could occur; looking beyond simplistic explanations for incidents); and containing problems (being prepared to deal with contingencies; using relevant expertise regardless of where it is situated within the organizational hierarchy). In their research on "high reliability organizations" they analyzed how highly regulated and standardized organizations such as nuclear plants, aircraft carriers and firefighting units achieve resilience in a complex environment. They found that technical checklists were not the key to success but instead a list of cultural features they define as "mindful organizing." They identified five principles that include preoccupation with failure, a reluctance to simplify, a sensitivity to operations, a commitment to resilience and a deference to expertise as a shared set of values that foster resilience through constant communication and recalibration in the face of unknowable risks (Weick & Sutcliffe 2015). High reliability theory should be explored in the cyber field because of the high potential for catastrophic outcomes from a cyber incidence. For instance, a cyber-attack on a nuclear plant could have long lasting and deadly consequences.

5.3 Improving the security of nuclear cyber security systems

Much has been written in the literature about nuclear security and safety in terms of nuclear reactors, waste disposal, site protection and security by design. However, little attention has been paid to the security of the cyber security system itself and the potential for catastrophic loss as magnified in the 2020 Solar Winds case in the United States where the country's largest multinationals and the U.S. government were caught off guard by an attack on their cyber security vendor (Constantin 2020). The Solar Winds cyber-attack illustrates the necessity of having a link between the supply chain and the organizations they support (Greiman and Bernardin 2021). Just as frightening as the scope of the attack is how smoothly the attack circumvented government and private-sector cyber defenses (Morris and Hackett 2021).

The application of 'security by design' in nuclear new builds will improve the protection of the plant and reduce the need for costly security improvements during its operating life. Security by design cannot fully protect a nuclear power plant from rapidly evolving cyberattacks, however, careful design of security systems and architecture can achieve levels of protection that exceed current norms and expectations (Brunt and Unal 2019). Security by design may well include a requirement for a technical support organization to conduct quality assurance of cyber defenses and practices, and this regime should be endorsed by a facility's executive board and continued at regular intervals after the new build facility has been commissioned. "Security by design cannot be a panacea, but it is an important factor in the establishment of a robust nuclear security – and cybersecurity – culture" (Brunt and Unal 2019).

6. Conclusion

Cyber security risk in nuclear power plants in the U.S. and around the world is a major national security threat impacting all nations. In light of the growing need for cleaner, low carbon technologies nuclear energy continues to evolve through the development of small modular reactors and newer fusion technologies. Countries seeking to develop nuclear energy capabilities should put in place the regulatory structures required to implement nuclear security, such as an independent regulatory body and a comprehensive set of regulations. Regulations should address the vital areas of insider threat prevention, cybersecurity, cyber resilience and a security culture, as well as traditional areas such as physical protection, control and accountability, and response capabilities. Moreover, national governments should take an active role in overseeing and centralizing at an institutional level the risks from cyber-attacks on nuclear facilities and the opportunities to mitigate these risks through more resilient structures. It is imperative that policy makers and private industry work together in partnership to develop more resilient security at the design, operational, systems and socio economic levels.

Acknowledgements

I would like to thank Renee Macasaet, my graduate research assistant for the collection and organization of extensive amounts of research for this paper.

References

Arinze, U.C., Longe, O.B., and Eneh, A. H.(2020) "Regulatory Perspective on Nuclear Cyber Security: The Fundamental Issues," International Journal of Nuclear Security Vol 6, No. 1, Article 3. https://doi.org/10.7290/ijns060103 Available at: https://trace.tennessee.edu/ijns/vol6/iss1/3

Atomic Energy Act of 1954, as Amended in NUREG-0980. Atomic Energy Act of 1954 (P.L. 83–703). https://www.nrc.gov/docs/ML1536/ML15364A497.pdf#page=23.

Brunt, R. and Unal, B. (2019) Briefing: Cybersecurity by Design in Civil Nuclear Power Plants. Chatham House, Royal Institute of International Affairs, International Security Department. London, UK.

Campbell, A. and Singh, V. (2019) "Lessons from the cyberattack on India's largest nuclear power plant" November 14, Bulletin of the Atomic Scientists.

Chaudhury, D.R. (2019, Aug 19) Kudankulam power plant 3rd unit moves towards operationalization ahead of PM's Russia trip. The Economic Times, Noida, Uttar Pradesh, India https://economictimes.indiatimes.com/industry/energy/power/govt-may-fund-unrealised-input-cost-of-gas-based-central-power-psus/articleshow/96892616.cms.

Constantin, L. (2020, December 15) SolarWinds attack explained: And why it was so hard to detect, IDG Communications. https://www.solar winds attack.

Cyber Security and Infrastructure Security Agency (2022) Shields-Up. CISA, Washington, D.C. https://www.cisa.gov/shields-up.

Department of Defense Science Board (DSB, 2017) Final Report of the Task Force on Cyber Deterrence, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, Washington, D.C., February.

Greiman, V. A. and Bernardin, E. (2021) Cyber Resilience: A Global Challenge. Academic Conferences and Publishing Ltd. (ACPIL), Reading, UK.

Holt, M. and Andrews, A. (2014) Nuclear Power Plant Security and Vulnerabilities. Congressional Research Service, Washington, D.C.

International Atomic Energy Agency (2021) IAEA Nuclear Security Series No. 42-G Implementation Guide: Computer Security for Nuclear Security, IAEA, Vienna Austria.

ITER Organization. (2022) What is Iter? https://www.iter.org/proj/inafewlines.

Kesler, B. (2001) "The Vulnerability of Nuclear Facilities to Cyber Attack," *Strategic Insights,* spring 2001, p. 15. http://www.nps.edu/Academics/Centers/CCC/Research-Publications/StrategicInsights/2011/Apr/SI-v10-i1_Kesler.pdf.

Kobayashi, T., Suzuki, T., and Iwata, K. (2019) Accident Cleanup Costs Rising to 35-80 Trillion Yen in 40 Years. Follow up Report of Public Financial Burden of the Fukushima Nuclear Accident. Japan Center for Economic Research (JCER).

Lloyd's. (2015) Lloyd's City Risk Index 2015-2025. Lloyd's, London, England.

Manto, C.L. and Lokmer, S. (eds.) (2015) Planning Resilience for High-Impact Threats to Critical Infrastructure, p. 199, Infraguard, Westphalia, Press.

Morris, B.Z. and Hackett, R. (2021, Jan 29) After SolarWinds: Untangling America's cybersecurity mess. Fortune, New York, NY.

National Energy Agency (NEA) (2020, June) Building low-carbon resilient electricity infrastructures with nuclear energy in the post-COVID-19 era. NEA Policy Brief, NEA, Organization for Economic Cooperation and Development (OECD), Paris, France.

National Institute of Standards and Technology (NIST) (2018), *Framework for Improving Critical Infrastructure Cybersecurity*, NIST, U.S. Department of Commerce, Washington, D.C. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

Nucnet (2020) Coronavirus: EDF implements Emergency Plan for French Nuclear Fleet, www.nucnet.org/news/edf-implements-emergencyplan-for-french-nuclear-fleet-3-5-2020.

Nuclear Energy Institute (2016) Cybersecurity for Nuclear Power Plants. NEI, Washington, D.C. Nuclear Energy Institute. (2016) NEI 08-09 Cyber Security Plan for Nuclear Power Reactors, Rev. 6. NEI, Washington, D.C.

Nuclear Threat Initiative (NTI) (2020) NTI Nuclear Security Index. NTI, Washington, D.C.

Nye, J.S. (2023, Jan 14) The Mouse Click that Roared. The Korea Times. https://www.koreatimes.co.kr/www/opinion/2022/02/197_142756.html

Nye, Jr., Joseph S. 2011. "Nuclear Lessons for Cyber Security" Strategic Studies Quarterly Vol 5, No. 4, pp. 18-38.

Office of the Director of National Intelligence (ODNI) (2021) Annual Threat Assessment of the U.S. Intelligence Community. U.S. Director of National Intelligence, 9 Apr. 20. https://www.dni.gov/files/ODNI/documents/assessments/ATA-2021-Unclassified-Report.pdf.

Organization for Economic Cooperation and Development (OECD) (2019) Resilience strategies and approaches to contain systemic threats, OECD Publishing, Paris.

Pickering, S. Y. and Davies, P. B. (2021) "Cyber Security of Nuclear Power Plants- US and Global Perspectives" Georgetown Journal of International Affairs*, online:* https://gjia.georgetown.edu/2021/01/22/cyber-security-of-nuclear-power-plants-us-and-global-perspectives/.

U.S. Department of Energy (2022) Cybersecurity Considerations for Distributed Energy Resources on the U.S. Electric Grid. USDOE, Washington, D.C.

U.S. Department of Justice (2022, Mar 24) Four Russian Government Employees Charged in Two Historical Hacking Campaigns Targeting Critical Infrastructure Worldwide. DOJ, Office of Public Affairs. https://www.justice.gov/opa/pr/four-russian-government-employees-charged-two-historical-hacking-campaigns-targeting-critical.

U.S. Nuclear Regulatory Commission (NRC) (2010) Cyber Security Programs for Nuclear Facilities Regulatory Guide 5.71, January. http://pbadupws.nrc.gov/docs/ML0903/ML090340159.pdf.

U.S. Nuclear Regulatory Commission (NRC) (2020) Report to Congress on the Security Inspection Program for Operating Commercial Power Reactors and Category I Fuel Cycle Facilities: Results and Status Update - Annual Report for Calendar Year 2020 (NUREG-1885, Revision 14). NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC) (2022) Draft Regulatory Guide DG-5061, Rev. 1 Proposed Revision 1 to Regulatory Guide 5.71. NRC, Washington, D.C.

Weick KE, Sutcliffe KM. (2015) Managing the Unexpected: Sustained Performance in a Complex World. (3rd ed.) John Wiley & Sons, Hoboken, NJ.

World Economic Forum (2021, Jan. 13) Nuclear: These countries are investing in small modular reactors. https://www.weforum.org/agenda/2021/01/buoyant-global-outlook-for-small-modular-reactors-2021.

World Institute for Nuclear Security (WINS) (2014) 4.3 Security of IT and IC systems at Nuclear Facilities, WINS International Best Practice Guide, https://wins.org/document/4-3-security-of-it-and-ic-systems-at-nuclear-facilities.

World Nuclear Association (WNA) (2023a) Nuclear Power in the world today. WNA, London, England. https://world-nuclear.org/information-library/current-and-future-generation/nuclear-power-in-the-world-today.aspx.

World Nuclear Association (2023b) Ukraine: Russia-Ukraine War and Nuclear Energy. WNA, London England.

World Nuclear Association (2022a) Chernobyl Accident 1986. WNA, London, England.

World Nuclear Association (WNA) (2022b) Fukushima Daiichi Accident (Updated May 2022) WNA, London, England.

Social-Engineering, Bio-economies, and Nation-State Ontological Security: A Commentary

Brandon Griffin[1-2], Keitavius Alexander[1-2], Lucas Potter[3], Xavier-Lewis Palmer[2-3]
[1]Old Dominion University, Norfolk, USA
[2]CySecSol, Franklin, USA
[3]BiosView, Oswego, USA

bgrif023@gmail.com
Kalex019@odu.edu
lpott005@odu.edu
xpalm001@odu.edu

Abstract: Biocybersecurity is an evolving discipline that aims to identify the gaps and risks associated with the convergence of Biology (the science of life and living organisms) and cybersecurity (the science, study, and theory of cyberspace and cybernetics) to protect the bioeconomy. The biological industries' increased reliance on digitization, automation, and computing power has resulted in benefits for the scientific community, it has simultaneously multiplied the risk factors associated with industrial espionage and the protection of data both commercial and proprietary. The sensitive and potentially destructive power of this data and its access inherently poses a risk to the national and ontological security of a nation. Ontological security refers to the extent to which an individual or group feels secure in their understanding of the world and their place in it. It is a psychological concept that pertains to the way in which people construct their sense of self and their place in the world, and how this sense of self and place is shaped by their interactions with others and the broader social, cultural, and political context in which they live. Nation-states provide stability and wider social cohesion, but these capacities can be disrupted when the nation state is sufficiently threatened (Bolton, 2021). Leading to an interest in maintaining a national identity; which can have profound effects on the behavior of a nation. Targeted social engineering is aimed at exploiting the changing and damaged mental health of workers in life science enterprises who have not been trained in a sufficient manner to deal with these attacks. Failure to identify the existing vulnerabilities associated with social engineering would expose the bioeconomy to unnecessary risk. Numerous scholars have pointed towards growing risks of nation-state stability being increasingly threatened vs inadequate actions taken to match threats for defense; when reflecting on energy, food, construction materials and more from the multi-trillion US bioeconomy we see that the ground to cover is huge (George 2019, Jordan, 2020, Murch, 2018; Mueller 2021). This paper seeks to discuss some of the existing vulnerabilities associated with social engineering attacks and the effects those attacks would have on the population's ontological security and spark conversations about ways in which ontological security of nation states are modified.

Keywords: Social-Engineering, Bio-economies, Nation-States, Bioveillance, Cyberbiosecurity, Biocybersecurity

1. Introduction

As the digital revolution continues to touch all walks of life the biological industry is no exception. The convergence of digitization, automation, and increased computing power has allowed for the rapid advancement of biological sciences while introducing new forms of risk to each sector of the bioeconomy. The acknowledgement of these threats and their intersections has allowed for the development of Cyberbiosecurity, which has been discussed at length (Murch 2018; Peccoud et al 2018; Murch and DiEuliis, 2019; Duncan et al, 2019; Richardson et al, 2019). Biocentricity of critical resources makes safeguarding the nation take on a different shape and conceptually this requires a more serious approach to STEM and its purveyors within and on behalf of the nation (Murch 2018; George, 2019). A particular threat that has the potential to disrupt both biological professionals and the general population is target social engineering attacks. These are attacks that rely heavily on human interaction and often involve manipulating people into breaking normal security procedures and best practices to gain unauthorized access to systems, networks or physical locations. The potential danger and unwanted consequences of these attacks would be catastrophic to the ontological security of a nation. These three concepts are interconnected and can have a mutually reinforcing effect on one another. Advances in biocybersecurity, for example, may be used to improve social engineering defenses, while social engineering tactics may be used to exploit vulnerabilities in biological systems. Disruptions to ontological security, meanwhile, may affect an individual's or group's ability to effectively defend against cyber threats. By studying these areas together, it is possible to gain a more comprehensive understanding of the complex interplay between them and to develop more effective approaches to addressing the challenges they pose.

2. Core Literature Considered:

Biocybersecurity literature has primarily focused on identifying the gaps associated with the ongoing digitization of the biological science industry. Research into biocybersecurity also seeks to provide recommendations to identify existing threats, and develop frameworks to anticipate, identify, mitigate and deter any hostile nation state or actors from participating in cyber biological attacks. One perceived gap is the lack of discussion surrounding ontological security in the midst of the cyberbio convergence of societies, especially those who are rapidly engaging 4th Industrial Revolution (4IR) technologies. Approaching biocybersecurity, social engineering, and ontological security as a commingled concept can facilitate the development of interdisciplinary approaches and solutions to these challenges. This may involve the collaboration of experts from diverse fields such as computer science, biology, psychology, and sociology, and may lead to the development of more effective and innovative solutions. By bringing together diverse perspectives and expertise, it is possible to gain a more comprehensive understanding of the challenges posed by these issues and to identify more effective approaches to addressing them. These technologies have the potential to more rapidly and specifically change the states of societies at larger scales than prior generation technologies and could use additional discussion. Tables 1 and 2 below respectively show the core literature considered for this work followed by pointing out selected foci in discussion that were helpful for driving this paper.

Table 1: Core Literature Considered

[1] Bolton, D. (2020) "Targeting Ontological Security: Information Warfare in the Modern Age", Political psychology, Vol 42, Issue 1, pp 127-142.	[6]Kavanagh, Jennifer and Michael D. Rich, Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life. Santa Monica, CA: RAND Corporation, 2018.
[2] Ramsey, F., & Seyyedhasani, H. (2021). Cyber attacks in agriculture: protecting your farm and small business with cyberbiosecurity.	[7] Mueller S. (2021). Facing the 2020 pandemic: What does cyberbiosecurity want us to know to safeguard the future?. Biosafety and health, 3(1), 11–21.
[3] Ferrara, E., Cresci, S. and Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. Journal of Computational Social Science, 3. doi:10.1007/s42001-020-00094-5.	[8] 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.
[4] George A. M. (2019). The National Security Implications of Cyberbiosecurity. Frontiers in bioengineering and biotechnology, 7, 51.	[9] Richardson, L. C., Connell, N. D., Lewis, S. M., Pauwels, E., & Murch, R. S. (2019). Cyberbiosecurity: A Call for Cooperation in a New Threat Landscape. Frontiers in bioengineering and biotechnology, 7, 99.
[5] Jordan, S. B., Fenn S. L. and Shannon B. B., "Transparency as Threat at the Intersection of Artificial Intelligence and Cyberbiosecurity," in Computer, vol. 53, no. 10, pp. 59-68, Oct. 2020	[10] G. Rowett, "The strategic need to understand online memes and modern information warfare theory," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4437-4442

Table 2: Selected Themes within Focused On

Papers	Selected Foci For Discussion
2, 4, 5, 8,	CyberBiosecurity (Defined)
1, 4, 5, 6, 9	Digitization of Society
5, 7	Social Engineering Risk
1, 2, 3, 4, 6	National & Ontological Security
1,3, 10	Informational Warfare & Social Media

3. National Roles in Ontological Security

The digital revolution has had a unique effect on the ontological security of nation states. Ontological security is defined as "a stable mental state derived from a sense of continuity regarding the events in one's life". Scholars discussing ontological security posit that individuals are not merely concerned with their physical security but also concepts such as their sense of being and perception of stability in their environment (Browning and Joenniemi, 2016). Historically, nation states have played a vital role in addressing this need, providing a stable environment and a national narrative that individuals are embedded within. As civil discourse continues to happen on social media the influence nations once had on the national narrative have effectively been decentralized. This decentralization has created attack vectors for hostile actors to create ontological insecurity through social engineering attacks targeted at the general public and the biological science industry. Throughout the years security professionals have worked to harden the physical systems to prevent unwanted intrusion. Despite being newly acknowledged as a crucial aspect of security engineering, the life sciences sector is not properly equipped to defend against psychological based attacks. Hostile nation states or lone actors are able to gain access to the system through the use of pretext, deception, and contact impersonations. Pretext is one of the quickest ways to bypass an enterprise's defenses, attackers targeting the bioscience fields often masquerade as official tools offering solutions to unique cyberbiosecurity challenges. Attackers lean into the demand for products and services such as research and bioinformatic tools to created complex social engineering campaigns that involve fake websites, and phishing scams that appear to come from official organizations such as HHS (Health and human services), CDC (Center for Disease Control) or WHO (World Health Organization) with the intention to spread harmful software to targeted systems. Attackers may also attempt to impersonate contacts within the target organization to gain access to sensitive data and information. Attackers using specially targeted social engineering messages and techniques seek to exploit these values to cause unwanted harm or intrusion. If the social engineering attack is successful, it would allow the attackers to demoralize, devalue, create distrust, and disenfranchise creators, companies, clients, and more in the biological sciences and biotechnological industries. A map of crucial considerations can be found in Figure 3, illustrating influences via social engineering.

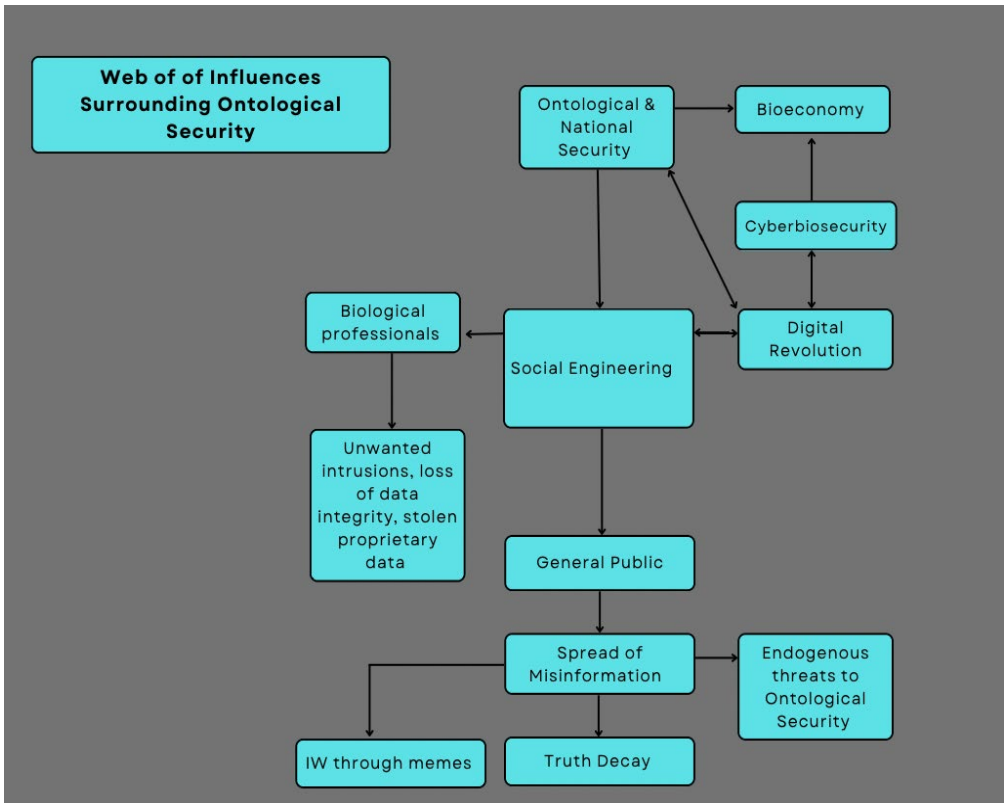


Figure 3: Concept Web of Influences Surrounding Ontological Security

Changes in information systems have allowed for hostile nation states to interfere in the domestic political conversation of their adversaries. The emergence of the internet and social media has amplified the reach of disinformation. Social engineering attacks targeted at the public look to cause ontological insecurity by disseminating truth decay within civil discourse. Truth decay can be defined with a set of four related trends that include: an increase of disagreement and misinterpretation of facts and data, inability to identify fact from opinion, opinion being valued over facts, and a growing in distrust toward respected sources of factual information. Some of the most prominent actors in this truth decay of the past 10 years can be observed within the transition from traditional media to internet and social media (Rich, 2018).

Truth decay helps democratize influence over the national narrative and poses an endogenous threat to ontological security. Endogenous threats occur when the national narrative out of which identity emerges appears disjointed from state behaviors, naturally this causes civil discourse on the topic. By distorting the information environment of the state, truth decay facilitates domestic questions of the congruence between policy and narrative. Cyber espionage can gather information, along with distorted or false information and images, which can be spread among a specific target audience in order to shape their perceptions of national narratives and particular policies or events. That is, leaked information can be distorted and bent towards biased and malicious aims before release to a target public. The possibility of cyber espionage in biological matters has implications for any leaks with the potential to shape national narratives (Bolton 2021). The effects of truth decay can be far reaching and destructive to a nation state, leading to the breakdown of civil discourse, gridlock in the political system, alienation from political and civil institutions and general ontological insecurity (Kavanagh and Rich, 2018). It is important to address truth decay in order to promote a healthy and functioning society.

Industries within the bioeconomy are utilizing and incorporating cyber linked technologies to perform and curate new technologies, perform analysis, log and examine data in real-time, mass produce biological goods, gain in-depth insight like never before, and other essential functions and activities to advance scientific capabilities and knowledge. An advantage of the integration of advanced computation implementation in these industries is the transformation of a more effective, accelerated, and articulated system in which crucial information is obtained and communicated over digital platforms. A key disadvantage is the capacity of attack vectors that cyber threats gaining to these sectors due the the dependence on cyber-linked platforms, devices, and databases. Current technologies of the bioeconomy are already being put to use to provide solutions to key issues such as global warming, agricultural shortages and medical treatments.In modern times, technology such as laptops and smartphones support the development and advancement of agricultural industries. Computing and automation capabilities that smart technology provides has caused major changes in food production (Ramsey and Seyyedhasani, 2021).

Industrial technological tools and biological science work in synchronization to assist processes, productions, analysis, transportation, and other necessary business functions across all sectors. Future technologies of the bioeconomy such as bioengineering could have many applications from cosmetology and to militant biological weaponry. Providing defense capabilities, methods, and resources for the bioeconomy is essential for current and future operation. This extends to considerations to civil engineering in urban planning with pandemics in mind (Potter et al, 2021).

Cyber attackers and adversaries have long since had means to exploit the cyber vulnerabilities in the bioeconomy and we can expect their means to deepen. Current cyber defense and countermeasures should be reassessed to account for ongoing threat evolutions for these systems, the technologies they produce, and the companies and agencies they belong to, and this includes a wider degree of training for personnel and infrastructural upgrade considerations (Duncan et al, 2020; Jordan, 2020; Elgabry et al., 2022). If hostile actors gain access to these core biotechnology intellectual property, a number of consequences could occur; specifically, information of the technology could be leaked to the public out of context or misleadingly with the purpose providing needless or inflated wariness. This in turn harms science communication initiatives and effective public health campaigns.

4. Social Media and Ontological Security

4.1 Role of social media in health promotion and important stakeholders

Social media is a tool in which people communicate on a digital platform sharing pictures, videos, and more digital content. Digital interaction between social media users can occur between different countries and continents. Content can communicate efficiently and quickly. The benefit of this is the creation of a community and ecosystem; professionals and nonprofessionals interact and share knowledge on social media platforms. Medical practitioners utilize social media interfaces to examine, communicate, treat, and triage patients. During critical events such as the COVID-19 pandemic, healthcare misinformation can be spread quickly (Adler et al, 2021; Mueller 2021; Palmer et al, 2021). This cyber access became a necessary component of healthcare in times while people were told to isolate themselves. In some circumstances, social media may be the only means of gaining key information and new detail regarding current healthcare events. Social media provides the platforms for sharing of experiences, education, policies, precautions, and other healthcare content. It is a tool in which healthcare professionals, companies, in industries can perform informational campaigns to better provide knowledge of resources, treatments, struggles, and awareness. New healthcare facilities in poverty stricken communities can be made aware of their proximity and presence and this implication can for non-traditional spaces as well. Overall, social media has given people the ability to communicate vital healthcare information and has proven to be an effective way to promote and increase the spread of translational health communication strategies. It allows users to share, use, and create critical health information and makes advocacy campaigns more efficient by providing a faster means of communication(Stellefson et al, 2020). How it is responsibly supported is vital.

4.2 Pandemic Responses Impacted by Social Media Misinformation

During the COVID-19 pandemic, misinformation was spread all over social media. From the beginning of the pandemic, ignorance of knowledge and details of the coronavirus was exploited, and this was sometimes considered to be through state actor influence and local cyber actors with malicious intent. Researchers at The Center for Countering Digital Hate have concluded that just twelve accounts known as the 'Disinformation Dozen' are responsible for the majority of COVID-19 misinformation across social networking platforms (Imran Ahmed, 2021). This had considerable basis for consideration in terms of asymmetric warfare considerations that malicious parties may invest in (Palmer et al, 2021). Misinformation campaigns against vaccination had a large impact across all social media platforms. This created hesitation, stress, confusion and debate on these vaccinations. Many social media users lacked professional experience, education, and media literacy to filter misinformation and get to necessary information that would be beneficial to their health. The mass of misinformation became such an issue that professionals even faced issues deciphering which content was authentic. Social media platforms had to tag and label which pieces of content were likely to contain misinformation. Some groups of people who lacked social media literacy or likely social usage altogether such as the elderly or technologically illiterate were more likely to be deceived by misinformation or not obtain valid information written by scientific, government, or educational institutions. Cyber threats could use this confusion to exploit non savvy social media users to buy supplements or believe in non verifiable medication such as ivermectin, an anti-parasitic medication rumored on social media to be an effective treatment against the coronavirus (Schellack, 2022). There is no current model to estimate its potential damage to public health (Schillinger, Chittamuru and Ramírez, 2020).Social media platforms attempted to filter out misinformation; however were still unsuccessful due to the large volume of it that still exists and botched means of official curation and messaging. A remaining challenge exists in the ability to support open and well researched dialogue among a public so narrowly trained.

4.3 On Practical Strategies for improving Community Health Literacy

Healthcare industries could create their own links and interfaces that interact with current and emerging social media platforms. Digital campaigns on social media platforms could be curated with the platforms themselves and creatives. Underserved communities, discriminated groups, and unliked social media users should be targeted to encourage engagement with these communities. Inclusivity can provide benefits in emphasis to

cover weak spots in messaging across cultures outside of mainstream habit. Given information to non professional social media users on how to differentiate misinformation from actual research. Social media should provide fundamental education on social media literacy. Relationships between social media platforms and professionals should be developed to create a cyber ecosystem that encourages health literacy. Rather than creating divided sides where content is politicized on social platforms, divided sides of social media users, dialogue discussing matters should be conducted to separate fact from fiction. Patterns, data, and trends of user activity may be analyzed to help healthcare education specialists and other professionals communicate ideologies. Data such as measurements of engagement and experience may help in making future decisions and preventing previous miscalculations and assessing progress of communicating information (Levac and O'Sullivan, 2010; Stellefson et al, 2020). Utilizing social media tools such as analytics can give healthcare professionals additional data to develop relationships and resources for social media platforms and users.

5. The Value of Securing Bioinformatics and databases to the bioeconomy

Emerging and existing technologies boost bioinformatics in examining biological significance and developing technological advances from them. The analytical assets and IP generated from them remain significant targets. A cyber attack from digital adversaries involving social engineering is especially dangerous to the development, reputation, and stability of companies in these industries. A successful attack involving social engineering and informational warfare in which key data on a new technology such as the generation of a cure or treatment through biodetection and bioengineering could result in loss of public trust. In recent times, the COVID-19 MRNA vaccines faced heavy scrutiny on social media due to the purposeful spread of misinformation. Memes, posts, videos, and more digital content were uploaded on many different social media platforms and observed by many. This caused distrust, division, and politicization of this vaccine. The emergence of harmful content on social media platforms that users can encounter in the online ecosystem, specifically regarding the COVID-19 pandemic, represents growing infodemics. This is the uncontrolled spread of information, including a multitude of low-credibility, fake, misleading, and unverified information (Ferrara, Cresci and Luceri, 2020.)

Similar events and circumstances can occur with evolving biotechnologies that utilize digitized bioinformatic systems to operate more efficiently and automated experimentation. Consequences and repercussions will be severe if a cyber threat gained access to a component of one of these systems and obtained sensitive information on the emerging biotechnology. Through means of social engineering and informational warfare, a malicious depiction of the technology could be shared to the world. Misinformation surrounding the technology could be used to sew mistrust in the general public which could have numerous consequences. These consequences include but are not limited to the general public's reluctance to embrace the new technology, the company or agency developing the technology losing financial and social capital, a slowing or ending of the scientific expedition, politicalization of the technology, theft and unauthorized replication of the intellectual property, a knowledgeable threat reverse engineering the development to formulate a weapon, and further cyber attack campaigns from threats to spread more falsehoods about the more technology from the same or similar scientific investigations.

6. Concluding Remarks, Including Limitations

Altogether, not identifying the risk potential of social engineering and informational warfare campaigns targeted at the biological science industry would cause tremendous damage to a nation's national and ontological security. Studying biocybersecurity, social engineering, and ontological security together can provide a more comprehensive understanding of the complex and interconnected nature of cyber threats and the ways in which they can impact individuals and society. By considering these issues in a holistic manner, it may be possible to identify and address potential vulnerabilities more effectively. For example, understanding the psychological mechanisms underlying social engineering tactics may help to develop more effective defenses against them, while understanding the impact of technological change on ontological security may help to mitigate potential negative effects. It is important to also think of how other mediums of relaying information and moods such as through music or video about information may play a role (Omar and DeQuan, 2020: Sice et al, 2020). The information gained from successful social engineering attacks is often used nefariously to question national narratives, inflame domestic policy conversations, and sow distrust among the population. Misinformation

spread through social media is often used by hostile actors to promote truth decay within target nations hoping to increase polarization within domestic politics that are intended to create political paralysis and unravel the bonds of society. There has been a push for biological industry professionals to build relationships and campaigns to help combat the spread of misinformation on the platforms. It is important that the biological industry continues to examine the risk factors associated with attacks that target human psychology. There has also been limited research and information on a systematic approach to educating the general public of the risk associated with social engineering and informational warfare campaigns that happen across social media platforms.

Finally, studying these three areas together can help to ensure that the impact of cyber threats on individuals and society is considered in the development of cybersecurity measures and policies. This can help to ensure that the interests of all stakeholders are considered and that solutions are effective and ethically sound. By considering the social and psychological dimensions of cyber threats, it is possible to develop solutions that are more sensitive to the needs and concerns of individuals and communities, and that take into account the broader societal implications of these issues.

References

Adler, A., Beal, J., Lancaster, M. and Wyschogrod, D., 2021. Cyberbiosecurity and Public Health in the Age of COVID-19. In Emerging Threats of Synthetic Biology and Biotechnology (pp. 103-115). Springer, Dordrecht.

Bolton, D. (2020) "Targeting Ontological Security: Information Warfare in the Modern Age", Political psychology, Vol 42, Issue 1, pp 127-142.

Browning, C. S., & Joenniemi, P. (2017). "Ontological security, self-articulation and the securitization of identity." Cooperation and conflict, 52(1), 31-47.

Carneiro, R., Dunca, S., Ramsey, F., Seyyedhasani, H. and Murch, R. (n.d.). "Cyber attacks in agriculture: protecting your farm and small business with cyberbiosecurity." [online] Available at: https://caia.cals.vt.edu/content/dam/caia_cals_vt_edu/cyberbiosecurity/2021%20Carneiro%20et%20al%20Cyberattacksinagricultureprotectingyourfarmandsmallbusinesswithcyberbiosecurity.pdf [Accessed 21 Oct. 2022].

Chen, J., & Wang, Y. (2021). "Social media use for health purposes: systematic review." Journal of medical Internet research, 23(5), e17917.

Duncan, S.E., Zhang, B., Thomason, W., Ellis, M., Meng, N., Stamper, M., Carneiro, R. and Drape, T., 2020. "Securing Data in Life Sciences—A Plant Food (Edamame) Systems Case Study." Frontiers in Sustainability, p.10.

Duncan, S.E., Reinhard, R., Williams, R.C., Ramsey, F., Thomason, W., Lee, K., Dudek, N., Mostaghimi, S., Colbert, E. and Murch, R., 2019. "Cyberbiosecurity: A new perspective on protecting US food and agricultural system." Frontiers in bioengineering and biotechnology, 7, p.63.

Elgabry, M., Nesbeth, D. and Johnson, S. (2022) "The Future of Biotechnology Crime: A parallel Delphi study with non-traditional experts," Futures, 141, p. 102970. Available at: https://doi.org/10.1016/j.futures.2022.102970.

Farsi, D. (2021). "Social media and health care, part I: literature review of social media use by health care providers." Journal of Medical Internet Research, 23(4), e23205.

Ferrara, E., Cresci, S. and Luceri, L. (2020). "Misinformation, manipulation, and abuse on social media in the era of COVID-19." Journal of Computational Social Science, 3. doi:10.1007/s42001-020-00094-5.

Gabarron, E., & Wynn, R. (2016). "Use of social media for sexual health promotion: a scoping review." Global health action, 9(1), 32193.

George A. M. (2019). "The National Security Implications of Cyberbiosecurity". Frontiers in bioengineering and biotechnology, 7, 51.

Jordan, S. B., Fenn S. L. and Shannon B. B., "Transparency as Threat at the Intersection of Artificial Intelligence and Cyberbiosecurity," in Computer, vol. 53, no. 10, pp. 59-68, Oct. 2020

Rich, M.D., (2018). "Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life." Rand Corporation.

Kavanagh, Jennifer and Michael D. Rich.(2018). "Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life." Santa Monica, CA: RAND Corporation .https://www.rand.org/pubs/research_reports/RR2314.html.

Levac, J. J., & O'Sullivan, T. (2010). "Social media and its use in health promotion." Interdisciplinary journal of health sciences, 1(1), 47-53.

Mwaura, J., Carter, V., & Kubheka, B. Z. (2020). "Social media health promotion in South Africa: Opportunities and challenges." African Journal of Primary Health Care and Family Medicine, 12(1), 1-7.

Mueller S. (2021). "Facing the 2020 pandemic: What does cyberbiosecurity want us to know to safeguard the future?." Biosafety and health, 3(1), 11–21.

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.

Murch, R. and DiEuliis, D., (2019). "Editorial: mapping the cyberbiosecurity enterprise." Front. Bioeng. Biotechnol, 7(235), pp.10-3389

Neiger, B. L., Thackeray, R., Van Wagenen, S. A., Hanson, C. L., West, J. H., Barnes, M. D., & Fagen, M. C. (2012). "Use of social media in health promotion: purposes, key performance indicators, and evaluation metrics." Health promotion practice, 13(2), 159-164.

Omar, B. and Dequan, W. (2020) "Watch, Share or Create: The Influence of Personality Traits and User Motivation on TikTok Mobile Video Usage", *International Journal of Interactive Mobile Technologies (iJIM)*, 14(04), pp. pp. 121–137. doi: 10.3991/ijim.v14i04.12429.

Palmer, X. L., Powell, E., & Potter, L. (2021). Matters of Biocybersecurity with Consideration to Propaganda Outlets and Biological Agents.

Palmer, X., Potter, L.N. and Karahan, S., (2021). "COVID-19 and biocybersecurity's increasing role on defending forward." International Journal of Cyber Warfare and Terrorism (IJCWT), 11(3), pp.15-29.

Peccoud, J., Gallegos, J.E., Murch, R., Buchholz, W.G. and Raman, S., (2018). "Cyberbiosecurity: from naive trust to risk awareness." Trends in biotechnology, 36(1), pp.4-7.

Potter, L., Powell, E., Ayala, O., & Palmer, X. L. (2021). "Urban Planning to Prevent Pandemics: Urban Design Implications of BiocyberSecurity (BCS)." In Intelligent Computing (pp. 1222-1235). Springer, Cham.

Ramanadhan, S., Mendez, S. R., Rao, M., & Viswanath, K. (2013). "Social media use by community-based organizations conducting health promotion: a content analysis." BMC public health, 13(1), 1-10.

Ramsey, F. and Seyyedhasani, H., (2021). "Cyber attacks in agriculture: protecting your farm and small business with cyberbiosecurity."

Richardson, L. C., Connell, N. D., Lewis, S. M., Pauwels, E., & Murch, R. S. (2019). C"yberbiosecurity: A Call for Cooperation in a New Threat Landscape." Frontiers in bioengineering and biotechnology, 7, 99.

Schillinger, D., Chittamuru, D., & Ramírez, A. S. (2020). "From "Infodemics" to Health Promotion: A Novel Framework for the Role of Social Media in Public Health." *American journal of public health*, *110*(9), 1393–1396. https://doi.org/10.2105/AJPH.2020.3057466.

Schellack, N., Strydom, M., Pepper, M.S., Herd, C.L., Hendricks, C.L., Bronkhorst, E., Meyer, J.C., Padayachee, N., Bangalee, V., Truter, I. and Ellero, A.A., (2022).0" Social Media and COVID-19—Perceptions and Public Deceptions of Ivermectin, Colchicine and Hydroxychloroquine: Lessons for Future Pandemics." Antibiotics, 11(4), p.445.

Sice, P., Elvin, G., Riachy, C., Shang, Y., Ogwu, S. and Zink, C., (2020). "Online screening of X-System music playlists using an Integrative Wellbeing Model informed by the theory of autopoiesis." IEEE Access, 8, pp.182307-182319.

Simpson, V. L., Hass, Z. J., Panchal, J., & McGowan, B. (2022). "Understanding the Development, Evaluation, and Sustainability of Community Health Networks Using Social Network Analysis: A Scoping Review." American Journal of Health Promotion, 36(2), 318-327.

Stellefson, M., Paige, S. R., Chaney, B. H., & Chaney, J. D. (2020). "Evolving role of social media in health promotion: updated responsibilities for health education specialists." International journal of environmental research and public health, 17(4), 1153.

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). "Prevalence of health misinformation on social media: systematic review." Journal of medical Internet research, 23(1), e17187.

Rowett, G. "The strategic need to understand online memes and modern information warfare theory," 2018

Search and CompAre Reverse (SCAR): A Bioinformatics–Inspired Methodology for Detecting File Remnants in Digital Forensics

George Grispos, William Mahoney, Sayonnha Mandal
School of Interdisciplinary Informatics, University of Nebraska-Omaha, USA

ggrispos@unomaha.edu
wmahoney@unomaha.edu
smandal@unomaha.edu

Abstract: A storage device may contain data that an individual is legally or morally not allowed to possess. Or, a disgruntled company employee may intentionally destroy corporate files, assuming once deleted the information is lost forever. The data could take the form of a database owned by a competitor, illegal images, or videos, or trade secrets or confidential business information. Fragments of the data may very well still be present on the disk drive, for example, and forensic tools may be capable of recovering some of the confidential information. This paper introduces Search and CompAre Reverse (SCAR), inspired from tools used in the bioinformatics community. The contribution is an initial empirical investigation into the use of this bioinformatics-inspired approach to deduce the partial existence of patterns in cases where traditional digital forensics tools cannot detect the type of the file due to overwriting the file signature portion of the file.

Keywords: Digital Forensics, Digital Investigations, Filesystems

1. Introduction

In the last decade, the digital forensics community has witnessed an increasing number and volume of storage devices being seized and examined by forensic investigators, which has created large and lengthy backlogs of investigations (Quick and Choo 2014). One of the reasons for these backlogs is the increased processing times needed to forensically image and analyse the storage devices under investigation. As a result, a growing number of digital forensics researchers (Richard III and Roussev 2006, Garfinkel 2010) have proposed that the community develop tools and techniques to decrease processing times needed to analyse individual devices.

Many current digital forensics tools have been designed to decode and present the underlying filesystem under investigation. For example, these tools can help recover files that have been marked as 'deleted', and then attempt to recover these files and present the results to the investigator. This same approach is used for both partially and 'totally' recovered files. Hence, these tools can be used by an investigator to provide evidence that a file or directory was once present on a storage device (e.g., hard disk drive), but has since been removed or deleted. However, a disadvantage to this approach is that in some cases when a file or directory is deleted, the 'space' occupied by this filesystem object could be reclaimed by the filesystem to store new files. The problem arises that many forensics tools require the detection and recovery of the file signature part of the file, to decode and present the file correctly to the investigator. Should the filesystem reclaim the signature portion of the file, these tools will be unable to decode the file signature and the remaining parts of the file could be marked as slack and grouped together with other slack files. Examining slack files can be time consuming, with an investigator effectively looking for a 'needle in the haystack'.

This paper attempts to solve this problem by introducing an alternative forensics methodology called *Search and CompAre Reverse (SCAR)*, inspired from previous research in the bioinformatics community. At a high-level, SCAR provides an investigator with a visual indication that a certain pattern (e.g., a file) either exists or once existed within a forensic image created from a storage device under investigation. The research contribution is an initial empirical investigation into the use of this approach to deduce the partial existence of patterns in cases where traditional digital forensics tools cannot detect the type of the file due to overwriting the file signature portion of the file. The SCAR approach can be used to determine if 'fragments' of a known file can be found on the blocks of a storage device under investigation, without relying on the underlying filesystem. The rest of this paper is structured as follows. Section two motivates the research, Section three describes the SCAR methodology, Section four presents the results of our initial analysis, and Section five provides results and their potential impact on the digital forensics community. Section six concludes the paper and presents ideas for future research.

2. Motivation

In short, digital forensics is the process of preserving and analysing digital evidence (Palmer 2001). A variety of digital forensics tools have been developed to assist investigators analyse storage devices associated with a specific digital crime or incident (Grispos and Bastola 2020, Freyhof et al. 2021, Grispos and Mahoney 2021). Many of these tools work by providing an investigator with a layer of abstraction between the bits stored on a storage device (e.g., a hard disk drive), and how this data is actually stored and organized on the storage device itself (Roussev et al. 2013). Common forensics tools used for the analysis of storage devices include Forensic Toolkit (FTK), EnCase, and Sleuth Toolkit.

These tools can be used to examine a specific filesystem and detect deleted files and directories, which is accomplished by the tool 'decoding' the filesystem details, and then searching for the specific file entries which are marked as deleted, but not yet overwritten. In some cases, data within the file (e.g., file signatures) may indicate the type of file (e.g., JPEG file). However, one issue with this approach is that it only works if the original filename entry and other file metadata can be recovered from the filesystem. This raises two questions: *what if the entry for a deleted file cannot be recovered? It is still possible to determine whether a certain file was stored on the filesystem, if we know what artifact we are looking for during our investigation?* While this answer can be found using many of the above tools, it is both cumbersome and time consuming.

While traditional forensics tools can be used to detect and recover deleted files, they usually require that the file header can be detected and decoded. Otherwise, the deleted file cannot be decoded or digested by these high-level abstraction tools. In cases where the file header has been overwritten, but the data segment of a deleted file is available, this data is often clustered together by the tools as 'slack files' (Casey 2009). These slack files can be very large in size and consist of unstructured data that have been recovered from 'orphaned' data segments that do not include file header information. Hence, the aim of this research is to develop a methodology that can be used to determine if the fragments of files, which could still exist in the filesystem of a storage device, can be used to determine, and identify known files. The nomenclature used here is to search for patterns within filesystem images; a pattern is simply the term used for a binary artifact such as a JPEG file. It must be noted that the actual filesystem type is not a factor in our approach.

A pattern can consist of an entire file, or a pattern can be a portion of a file. For example, if an investigator suspects that a filesystem contains an illegal JPEG picture, the investigator might use only a fraction of the JPEG picture as a pattern; in this manner the investigator does not need to have the original picture but can still decide as to whether that portion of the image is present on a filesystem. The aim of SCAR is to perform partial matches of blocks of bytes within the patterns against blocks of bytes within the filesystem (in this example, a JPEG picture within a filesystem). Matching partial blocks is critical because the patterns may be only partially resident on an image; the pattern or file may have been deleted and subsequently overwritten by the contents of new files.

This partial overwriting of blocks is quite common. Files are stored in blocks on the device, assigned by the operating system, and tracked as part of the filesystem. When a file is deleted, the data portion of the file is not overwritten; simply the link from the name of the file to the data for the file is removed, and the blocks allocated to the file are marked as 'available'. As files are created, these free blocks will eventually contain new data. Deleting a file and never creating any new files will leave the data intact for the life of the device. But if new data is created on the device, these blocks containing the previous information will gradually be eliminated. Moreover, if newer files are smaller than the original block size, older data can remain in the original block allocations. Hence, SCAR scans blocks from the back towards the front in order to identify older deleted files in larger blocks.

3. Approach

SCAR is influenced by previous work from the bioinformatics community since the problem being addressed could be very similar. While bioinformatics software is used to scan for the presence of a particular gene in a sequence of DNA, a forensics investigator might design software to scan for the presence of a file in a storage device.

3.1 Prior Bioinformatics-Inspired Cybersecurity Efforts

In the past, cybersecurity researchers have attempted to leverage the similarities in searching for a specific gene in a digitised sequence of DNA, with searching for digital artifacts in filesystems. Much of this research has expanded the ideas and approaches proposed by Altschul et al (1990) for the Basic Local Alignment Search Tool

(BLAST), a computer algorithm that can be used online (National Center for Biotechnology Information n.d.), or installed on a local machine. BLAST is used to rapidly align and compare a query DNA sequence with a database of other, longer sequences, to find 'alignments' (i.e., approximate matches) that can be used a starting point for further investigation.

Within cybersecurity, BLAST has been used to classify new malware (Pedersen et al. 2012, Pedersen et al. 2013), where it was used to search for similarities between sequences of a digital artifact (a virus), and a set of sequences representing other digital artifacts, which are stored in a BLAST database. Computer virus code was converted into the representation necessary for the tool (e.g., search strings), which were then compared using BLAST to previous collections of computer viruses, to classify the new malware.

Fink and Oehmen (2012) developed an approach that extends the Machine Learning String Tools for Operational and Network Security (MLSTONES) project active at Pacific Northwest National Labs. They demonstrated that their approach could be used to convert digital binary files into amino acid sequences, which are then matched against the MLSTONES system.

The CodeDNA project (Glendenning n.d.) uses a longer alphabet to represent executable code. Sections of software are assigned a 'letter' depending on the function, and the functions are higher-level. One code might indicate a file operation, while another indicates creating a socket connection. Once potential malware is analysed and then translated into a sequence, the CodeDNA software can then be used to compare the sequence to the DNA of known malware to identify resemblances.

Other uses for DNA analysis in cybersecurity include encryption (Priya and Saritha 2017), using sequence alignment for network threat recognition (Kozakiewicz et al. 2007), and using pair-wise sequence alignment to test for masquerading (Coull et al. 2003). Hence, while previous research has examined the use of BLAST for various applications in the cybersecurity domain, this tool could also, in theory, be used to scan an entire filesystem for remnants of known files for use within digital forensics. However, one challenge that could emerge is the that one must convert the filesystem and the known file pattern into DNA sequences in ASCII, and only then provide these DNA sequences to the tool.

3.2 Description of SCAR

Our SCAR methodology is an expansion of concepts presented by Neamatollahi et al. (2020). The approach presented designed towards an exact match of a subsequence within a larger DNA database. The authors present a set of algorithms initially scan the DNA database to create 'windows' where there is a potential match with the search sequence. These 'windows' represent starting points for subsequence matches, which are then examined to a greater extent for the exact matching.

One of the algorithms presented by Neamatollahi et. al. is the Processor-Aware Pattern Matching (PAPM) algorithm. This algorithm is considered processor-aware, since rather than examining and comparing DNA byte by byte, it compares several bytes at once; a 64-bit CPU will examine eight ASCII characters. The algorithm, unfortunately, assumes that the CPU allows unaligned 64-bit memory accesses, so while this would execute on x86-64, it will not work on many reduced instruction set architectures.

In the case of digital forensics, PAPM is a close fit, but with the allowance that the search proceeds in opposite direction. The aim in this case is to determine whether file pattern fragments still exist on the filesystem. These pattern fragments may be partially overwritten in the disk block, and if so, the section of the disk block that would have been overwritten is at the start of the block, with the original fragment located at the back. Thus, a portion of the SCAR algorithm is to search the disk blocks in the same manner as PAPM but from the end of the block towards the front. There is only one starting point window to consider, and that is the last CPU word in the block under test. Therefore, the search comparison proceeds from the last CPU 'word' of the block towards the front of the block, and if a word-size mismatch occurs, the last trailing bytes are tested so that an exact number can be returned. SCAR can simplify DNA subsequence searching, as we are only interested in exact matches on partial blocks, as opposed to bioinformatics searches, where the subsequence may have minor differences.

3.3 SCAR Algorithm

Algorithm 1 provides a broad overview of the methodology used by the approach. The algorithm scans for file contents (patterns) within an acquired forensic image. In Algorithm 1, "image blocks" are large sections of a filesystem, while "pattern blocks" are smaller sections of what is currently being searched for. The algorithm

maintains a “search set” which is the information necessary to look for one file within the image blocks. Search set entries are treated as a pool of resources and are assigned to pattern files as needed. Each search set entry has an indicator which tracks the status of the entry, whether the entry needs the next block of pattern data, whether it is ready for using the CPU, and so on. All search set entries have a status of “available” when the algorithm starts.

```
while (keep_going)
       Reset the image file input to the beginning of file

       while (keep going and more image blocks to read)

              for (each entry in the search set)
                     if (this entry is “completed”)
                            Print the results
                            Mark the entry as “available”

              for (each entry in the search set)
                     if (this entry is “available”)
                            Open the next pattern and assign to this entry
                            Mark the entry as “needs_data”

              for (each entry in the search set)
                     if (this entry is “needs_data”)
                            Read the next block from the pattern
                            Mark the entry as “needs_CPU”

              for (each entry in the search set)
                     if (this entry is “needs_cpu”)
                            Perform the PAPM right-to-left search

              for (each entry in the search set)
                     if (this entry is “needs_cpu”)
                            This entry has scanned all blocks in the image
                            Mark the entry as “needs_data”

              for (each entry in the search set)
                     if (all entries are “available”)
                            We have processed all patterns
                            keep_going = false
```

Algorithm 1: Primary SCAR Program

The motivation for this algorithm is the consideration that, potentially, an investigator could be searching for different patterns when searching a storage device, and that these patterns are not of a uniform size. As a result, smaller pattern files will result in a quicker search as compared to larger pattern files, making entries in the ‘search set’. available for use by the next pattern file. Further, the program is trivially parallelisable. This is an important requirement as one assumes that the input (a forensic image) is likely to be the size of a filesystem. The overall algorithm is designed to take advantage of the POSIX thread library available on Linux and other systems. Tuning the size of the search set to the number of CPU cores available maximizes the utilization of the processors; looking for ten patterns at a time would cost little more in terms of time than looking for five, assuming there are sufficient CPUs. Moreover, after the threads are started and performing their right-to-left search of the pattern blocks, the main loop of the program initiates a read for the next block of the image. Observation shows that in general this allows the next image block to be available by the time the threads have completed the search through the current dataset, thus overlapping the I/O operation.

The overall flow of the method is to: a) create as many search set entries as necessary to match the CPU cores, b) load in as many patterns as there are search set entries, and c) for each entry compare the pattern blocks against the current image block, for all pattern and image blocks currently available. This is repeated until the entire forensic image is scanned, while filling in search set entries with new patterns as the entries become available. It is necessary to have larger patterns scan through the image more than one time, and this is time consuming. From an optimization perspective, if a pattern (e.g., a JPEG file) has been completely matched, there is no need to continue searching for the pattern blocks. Thus, a pattern that is deleted from the filesystem, but not overwritten can cause the search to finish for that pattern.

```
// t is the address in memory of the image block,
// p is the address in memory of the pattern block
match_count = 0
t_ptr = t + size of the image block - size of a word
p_ptr = p + sized of the pattern block – size of a word
while (t_ptr > t and p_ptr > p and word at p_ptr == word at t_ptr)
       match_count = match_count + size of a word
       t_ptr = t_ptr – size of a word
       p_ptr = p_ptr – size of a word
```

Algorithm 2: Processor Aware Pattern Matching

What remains is the description of the PAPM right-to-left-search, based on ideas proposed by Neamatollahi et. al., (2020) and shown above in Algorithm 2. This algorithm progresses right to left within the block, and counts matches in terms of 8-byte (64-bit) quantities instead of bytes.

3.4 Performance Parameters

As this is an experimental algorithm, the parameters for the algorithm should be changeable. In terms of performance, the algorithm is influenced by several different factors:

- **The size of the filesystem image block currently in memory**: There will be a certain amount of time required to scan for the pattern blocks within the image blocks, and larger image blocks will of course take proportionally longer to scan. But the trade-off is that it is possibly faster to read larger amounts of the image into memory in one input request to the operating system.
- **The size of the pattern block currently in memory**: The same question exists for the pattern blocks. Suppose one reads in 64KB of a JPEG file, but that the file is 65KB in size. With the current algorithm, the remaining 1KB of data will need to be compared against the entire filesystem, necessitating a complete pass through the filesystem for only a small portion of the pattern. Here one can assume that 'bigger is better' and for this reason allocating 128KB would perform better. But it is not essential to have the entire original file as the pattern; one could take a 64KB sample from the middle (aligned on a block) of the file and use this as the pattern. The files used in the evaluation were all less than 64KB in size, so in terms of performance this parameter was not tested.
- **The number of CPU cores available for searching**: Here the trade-off is that if the hardware has 16 cores, one can perform 16 pattern searches simultaneously. The algorithm operates correctly whether the number of threads is smaller than, equal to, or larger than the number of cores. The trade-off is the time required to start the threads and to rendezvous with their completion.

It is necessary to select parameters for SCAR that are appropriate for the user's hardware. For example, it may not make sense to execute 24 threads on a machine with two cores. To evaluate the performance of the algorithm, the authors explored some of the above issues, using a HP ProLiant with 160 GB and 24 cores. The following results were observed. Adding more threads improves performance, but only up to a certain point. Various Linux tools such as *htop* and *strace* show that most pattern-block-to-image-block matching fail quickly; the overhead of starting threads, which then fail, tends to consume more time than the payback of testing the patterns in parallel. A thread limit between 4 and 10 seems appropriate on this hardware. This leads to the question: is it better to read smaller or larger blocks of the patterns and the image? To determine this impact, the size of the in-memory block from the filesystem image was varied, since it is the larger of the two variables. On the HP hardware, the time required to read filesystem images in blocks over 8MB will have an impact on the total elapsed time required to perform the tests. The threads running the PAPM test will complete and then wait before the next filesystem image block is available. Using the Linux command *htop*, it quickly becomes apparent that a very large setting for the size of the image blocks means that the available CPUs are not being utilized 100%. Rather, threads start and immediately complete, and the algorithm waits for the next block.

It is advantageous to run the algorithm with a few samples to determine good settings for specific a hardware setup. SCAR allows a user to set from the command line, the number of threads, the size of each pattern block, and the size of the image blocks, as well as the file containing the image and the directory containing the patterns.

4. Forensic Usability of SCAR

SCAR was evaluated to determine the usability and effectiveness of using the approach to locate patterns (files) in a filesystem, during a forensic investigation. The test environment consisted of a virtual machine running Ubuntu Linux version 20, setup with four partitions. The primary partition was used to store the operating system and to execute SCAR, while the three other partitions (hereafter referred to as the 'test partitions') consisted of a FAT32 filesystem, approximately 64MB in size. The test partitions were filled with JPEG image files ('the test dataset') from data sets of images of dogs (Khosla et al. 2011) and flowers (Nilsback and Zisserman 2008), in order to represent 'pattern files. These JPEG images were selected at random, and the same random selection were then copied onto both test partitions. In cases where random parts of a pattern file are overwritten, the Linux device */dev/urandom* was used as the source of the random bytes. The test dataset on the test partitions were then manipulated in three different tests as discussed below. At a high-level, these manipulations involved a treatment (deletion and/or overwriting of files) being applied to the test dataset, and then the analysis of the test partitions using SCAR methodology and Autopsy, a toolkit commonly used by the digital forensics community. Ten files were randomly selected from the test dataset to be used as patterns for SCAR searches in all three tests.

4.1 Detection of Deleted Files

The first test involved the detection of deleted files within the first test partition. This involves checking that the algorithm operates correctly, and that Autopsy detects the deleted files as a confirmation. After mounting the test partition, ten random files are deleted from the filesystem. The filesystem is then unmounted and subjected to both the SCAR methodology and Autopsy. With respect to SCAR, the algorithm detects 100% of the pattern blocks for all ten files. Likewise, when the partition was examined using Autopsy, all ten files were recovered and presented for analysis. In both cases, the results are not unexpected since the contents of the files have not been destroyed or overwritten. Hence, both SCAR and Autopsy can 'see' the deleted files.

4.2 Deleted Files with Random Blocks Partially Overwritten

The purpose of second test was to investigate if SCAR can be used to detect partially deleted files from the test partitions. For this test it is assumed that the blocks, which are no longer part of file, are assigned in some order, either by disk cylinder or by a closest-seek-time-next algorithm that is not related to the ordering of the blocks in the file. In these cases, blocks of the file will be destroyed in random order. Using the second test partition, ten random files were deleted from the filesystem: 25% of the blocks in each file were selected at random, and of those blocks, the first 25% of the block was overwritten with random data. It can be assumed that for some of these files, the file headers could still be present on the filesystem, with others may have been destroyed. This sets a scenario where files are deleted, and the filesystem reallocates the space to 'new' files. The analysis using Autopsy revealed that all ten files could be detected, and an attempt was made to recover the data within these files. However, only one of these ten files could be visualised using the file viewer in the toolkit. Further analysis of this file revealed that it was the only file whose file header was still intact, allowing Autopsy to visualise the file to the user. Figure 1 below presents the original file from the test data set, and the file recovered using Autopsy with 25% of the file randomly 'deleted'. The test was also repeated by overwriting 50% of the blocks in the ten random files, as well as 75% of the blocks in the same files. In both cases Autopsy did not detect any of the files.



Figure 1: Original JPEG and JPEG Image with 25% Data Destroyed

The test partition was then analysed using SCAR and visualisations were created to show the file patterns visible on the partition. For example, Figure 2 presents the pattern file for the JPEG image shown in Figure 1, with 25% random blocks destroyed, 50% destroyed, and 75% destroyed. To visualize the pattern data, we use a scale colour-coded by the percentage of the pattern block that was discovered intact on the partition. For this specific pattern, where there happen to be 52 data blocks, the visualization has eight pattern blocks horizontally, so the last row is only partially filled.

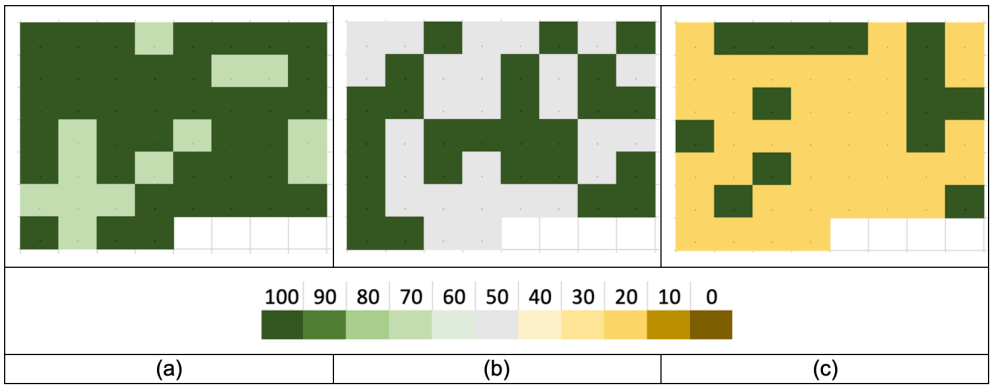


Figure 2: **SCAR Detection with 25% Random Data Destroyed (a), 50% (b), and 75% (c)**

The visualisations show that SCAR can still detect and confirm the remnants of the original file, even though the actual file is missing significant amounts of data. The SCAR output also shows why at 50% and 75% destruction the file cannot be rendered by Autopsy as the file headers are destroyed but can still be detected and confirmed by SCAR.

4.3 Deleted Files with Complete File Partially Overwritten

When a file is deleted, the blocks belonging to the file are available for other use, but different filesystems handle the free blocks in different ways. Blocks could be allocated in a first-in-first out order, or possibly selected by their proximity to other blocks in the same file. To create a realistic test scenario to mimic this situation, ten random files were chosen and 25% of the file content was overwritten with random data. The file was then deleted. These actions simulate the file being deleted and then the data blocks being reused in order. The test was repeated with 50% and 75% overwritten.

The analysis of the test partition with Autopsy revealed that the tool detects the deleted file contents, but the files are not recognizable as pictures. However, Autopsy only detects the file because the filesystem directory entry had not yet been re-used. If the entry is overwritten, Autopsy would not discover the existence of the pattern data on the filesystem image. As a result, this file could go undetected. The analysis of the third partition with SCAR (as shown in Figure 3) shows that in all three cases (25%, 50% and 75% overwritten) SCAR can still detect remnants of the file. The argument can then be made that that a percentage of the file pattern still exists on the filesystem, even if portions of the file have been overwritten and the entry in the directory for the file is destroyed.

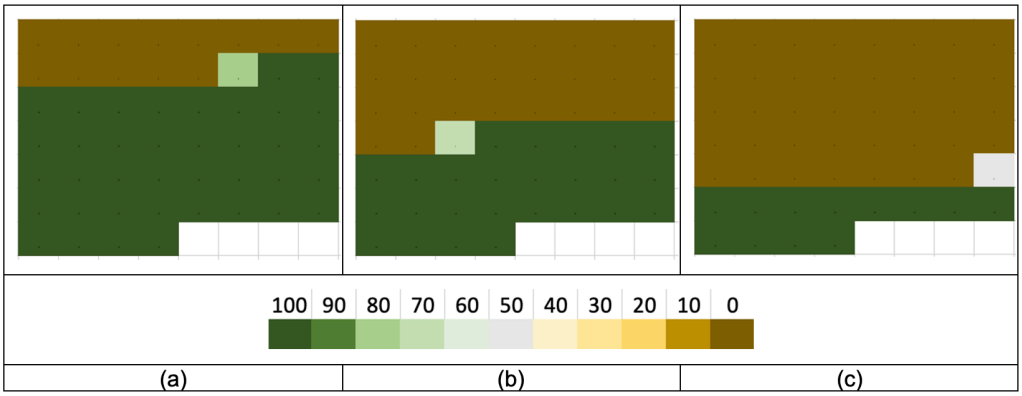


Figure 3: **SCAR Detection with 25% Initial Data Destroyed (a), 50% (b), and 75% (c)**

5. Discussion

A limitation to our approach is that the forensics investigator must have a copy of the original file to search for it in the investigation. Likewise, if the file in question changes frequently (e.g., a database) our approach might be limited as the investigator may have a copy of the file which is outdated relative to what is on the media being investigated. While these limitations are evident, our approach also has advantages over other tools.

For example, forensic investigations of illegal images usually proceed by creating a hash value for the files on a suspect disk and comparing these hash values to those of known illegal images. Files which have been deleted, but can be recovered, can be compared as well. But simply changing one byte of an image file will change the resulting hash, and as a result the file may be present and the image still able to be viewed, but the file goes unnoticed by the investigator. Similarly, if portions of a file have been overwritten, a different hash comparison value will be presented to the investigator.

If the investigator has a complete copy of what they are looking for, the SCAR algorithm will search for the matching data. But there are understandable legal and ethical ramifications if the investigator has an exact copy of an illegal image. In contrast to simply having a hash of the file, or the entire file itself, the investigator could possess a portion of the data from original file; not enough for them to visualize the actual photo, but enough for them to conduct a search of the suspects disk. If there is a collection of several illegal files, the investigator could maintain a collection of portions of the files. For example, they could maintain a set of 32 Kbyte samples of each of the files, starting 4 Kbytes into the data. This skips past file header information and directly into the image, so it cannot be visualized on her computer. But the 32 Kbytes of file data can still be searched for on the suspects media. And since our algorithm operates on blocks of 512 bytes, any offset which is a multiple of 512 is acceptable. The advantage of SCAR, then, is in instances such as these; the hash method may not be applicable for one of several reasons, including a partially deleted file or a slightly modified file. But if a determination can be made that the data from the file is still present on the suspect filesystem, this still provides some evidence of the prior activity. Whether that evidence is 'sufficient' for a particular crime is considered out of scope for this research.

6. Conclusions and Future Work

This paper presents SCAR, a digital forensics methodology that is inspired from concepts initially introduced by the bioinformatics community. At a high-level, the purpose of SCAR is to scan storage media, looking for fragments of a known file that could exist in the filesystem of the storage device. This can be useful from an intelligence perspective; a business employee may have copied confidential corporate information that they should not have in their possession, or the employee may have deleted files intentionally through malice. Since many filesystems overwrite deleted files quickly after the user has deleted a particular file, SCAR can be used to search the filesystem of the storage media under investigation, providing information to the investigator if a portion of the file can still be found in the filesystem. The positive identification of the 'portion' of the file could provide an investigator with enough information to show a 'smoking gun' that the filesystem did at some point contain the known file. In this sense, SCAR could be particularly useful in the identification of information in investigations concerning illegal pictures and videos, corporate secrets, intellectual property, and other business assets, although we do not address the question of 'how much' evidence is 'enough' evidence.

There are several avenues for future work. First, we intend to refine the SCAR development and evaluate the methodology against a larger number of file types. While this initial research focused on the identification of JPEG files, future research intends to focus on other image and video file types, which are commonly sought in other digital forensic investigations. Second, future work will investigate the effectiveness and refinement of the SCAR methodology in other filesystems. The idea behind this avenue of research is to determine if SCAR can be successfully executed on filesystems such as NTFS, and what changes (if any) need to be implemented to allow the approach to be used in these filesystems. Finally, future research intends to examine the applicability of the SCAR methodology in RAID-based storage, commonly used in server-based storage solutions. Since RAID can 'split' a file over more than one physical storage device, this potential research will examine how SCAR can be modified to help a forensic investigator identify the existence of a particular file on physical storage devices used in a RAID configuration. As discussed in Section 3.4, there are several "tunable parameters" involved in optimizing the algorithm on a computer system, and future work could include an automated method for adjusting these parameters for optimal performance.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Casey, E. (2009). Handbook of Digital Forensics and Investigation, Academic Press.

Coull, S., J. Branch, B. Szymanski and E. Breimer (2003). Intrusion Detection: A Bioinformatics Approach. 19th Annual Computer Security Applications Conference, 2003. Proceedings., IEEE.

Fink, G. A. and C. S. Oehmen (2012). Final Report for Bio-inspired Approaches to Moving-Target Defense Srategies, Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

Freyhof, M., Grispos, G., Pitla, S. and Stolle, C. (2022). Towards a Cybersecurity Testbed for Agricultural Vehicles and Environments. Proceedings of the Seventeenth Midwest Association for Information Systems Conference, Omaha, Nebraska.

Garfinkel, S. L. (2010). Digital Forensics Research: The Next 10 years. Digital Investigation 7: S64-S73.

Glendenning, M. (n.d.). CodeDNA: Scalable, High-Spoeed, High-Volume, Shareable Malware Detection. Cyber Security Division Transition to Practice Technology Guide, Department of Homeland Security Science and Technology Directorate: Cyber Security Division.

Grispos, G. and Bastola, K. (2020). Cyber Autopsies: The Integration of Digital Forensics Into Medical Contexts. 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE.

Grispos, G., and Mahoney, W.R. (2022). Cyber Pirates Ahoy! An Analysis of Cybersecurity Challenges in the Shipping Industry. Journal of Information Warfare, Vol. 21 (3), pp. 59-73.

Khosla, A., N. Jayadevaprakash, B. Yao and F.-F. Li (2011). Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs. Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Citeseer.

Kozakiewicz, A., A. Felkner, P. Kijowski and T. J. Kruk (2007). Application of Bioinformatics Methods to Recognition of Network Threats. Journal of Telecommunications and Information technology: 23-27.

National Center for Biotechnology Information. (n.d.). Basic Local Alignment Search Tool. from https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Neamatollahi, P., M. Hadi and M. Naghibzadeh (2020). Simple and Efficient Pattern Matching Algorithms for Biological Sequences. IEEE Access 8: 23838-23846.

Nilsback, M.-E. and A. Zisserman (2008). Automated Flower Classification Over a Large Number of Classes. 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE.

Palmer, G. (2001). A Road Map for Digital Forensics Research-Report from the First Digital Forensics Research Workshop (DFRWS). Utica, New York.

Pedersen, J., D. Bastola, K. Dick, R. Gandhi and W. Mahoney (2012). Blast Your Way through Malware analysis Assisted by Bioinformatics tools. Proceedings of the International Conference on Security and Management (SAM).

Pedersen, J., D. Bastola, K. Dick, R. Gandhi and W. Mahoney (2013). Fingerprinting Malware using Bioinformatics Tools Building a Classifier for the Zeus Virus. The 2013 International Conference on Security & Management (SAM2013).

Priya, S. K. and S. Saritha (2017). A Robust Technique to Generate Unique Code DNA Sequence. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE.

Quick, D. and K.-K. R. Choo (2014). Impacts of Increasing Volume of Digital Forensic Data: A Survey and Future Research Challenges. Digital Investigation 11(4): 273-294.

Richard III, G. G. and V. Roussev (2006). Next-Generation Digital Forensics. Communications of the ACM 49(2): 76-80.

Roussev, V., C. Quates and R. Martell (2013). Real-Time Digital Forensics and Triage. Digital Investigation 10(2): 158-167.

Development and Analysis of a Reconnaissance-Technique Knowledge Graph

Elsa Deitz[1], Eve Cohen[1], Jordana Wilkes[1] and Thomas Heverin[2]
[1]The Baldwin School, Bryn Mawr, United States of America
[2]Girls Learn Cyber, Inc., Philadelphia, United States of America

hacker@girlslearncyber.com
thomas@girlslearncyber.com

Abstract:Penetration testing involves the use of many tools and techniques. The first stage of penetration testing involves conducting reconnaissance on a target organization. In the reconnaissance phase, adversaries use tools to find network data, people data, company/organization data, and attack data to generate a risk assessment about a target to determine where initial weaknesses may be. Although a small number of tools can be used to conduct many of reconnaissance tasks, including Shodan, Nmap, Recon-ng, Maltego, Metasploit, Google and more, each tool holds an abundance of specific techniques that can be used. Furthermore, each technique uses unique syntax. For example, Nmap holds over 600 scripts that make up its Nmap Scripting Engine. Depending on the type of device targeted, Nmap scripts can scan for ports, operating systems, IP addresses, hostnames and more. As another example, Maltego operates over 150 transforms or modules that collect data on organizations, files and people. Understanding which reconnaissance tool, techniques within those tools, and the syntax for each technique represents a highly complex task. MITRE ATT&CK, a widely accepted framework, models tactics and techniques within the tactics to help users make sense of adversarial behaviours. The tactic of reconnaissance is modelled in ATT&CK as well as its techniques. However, the explicit links between reconnaissance techniques are not modelled. Our research focused on the development of an ontology called Recontology to model the domain of reconnaissance. Recontology was then used to form Reconnaissance-Technique Graph (RT-Graph) to model 102 reconnaissance techniques and the directional links between the techniques. We used exploratory data analysis (EDA) methods including a graph spatial-layout algorithm and several graph-statistical algorithms to examine RT-Graph. We also used EDA to find critical techniques within the graph. Patterns across the results are discussed as well as implications for real-world uses of RT-Graph.

Keywords: Knowledge Graph, Ontology, Reconnaissance

1. Introduction

MITRE ATT&CK serves as a model of adversary tactics and techniques that come from real-world observations. It represents a widely accepted knowledge base of adversary behaviours. ATT&CK has been widely accepted in both research and industry domains and has been applied to adversary emulation, red teaming, security tool development, and security operations center (SOC) assessments (Georgiadou, Mouzakitis, and Askounis, 2021). ATT&CK consists of 14 tactics that include Reconnaissance, Initial Access, Execution, Command and Control, Exfiltration, and others. Each tactic contains various techniques that adversaries use to achieve the tactic's goal.

We focused our research on the Reconnaissance Tactic of ATT&CK. MITRE (2022) defines reconnaissance as "…techniques that involve adversaries actively or passively gathering information that can be used to support targeting." Reconnaissance techniques (RTs) are used to collect data such as device names, IP addresses, host names, employee names, email addresses, and more. MITRE (2022) stated that information collected from one RT can be used "to drive and lead further reconnaissance techniques." Although ATT&CK conceptualizes that one RT can inform or lead to other RTs, ATT&CK does not explicitly model these technique-to-technique connections. Our research aimed to show how a knowledge graph could model direct connections among RTs and to show how graph statistical algorithms could be used to explore the data within the knowledge graph. Results and insights are provided about the graph statistical analyses run over the knowledge graph.

2. Background

The reconnaissance domain represents a difficult domain to understand due to various challenges. These challenges include navigating an abundance of available tools, figuring out the varied uses of each tool, sifting through various types of data produced by the tools, and understanding the relationships among all the data (Steinberg, 2017). Ontologies serve as one way to help reduce the complexity of these challenges. Ontologies model domain concepts and relationships among the concepts. Gruber (1993, p. 1) defined an ontology as "…a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects." People can gain an understanding of a domain from ontology. Ontologies can be used to create a body of knowledge about a domain. For example, Aviad, Wecel, Abramowicz (2015) used an ontology to model the cybersecurity body of knowledge.

An ontology's modelled concepts (also called classes) provide "data slots" for specific objects from a domain. For example, in the ATT&CK Reconnaissance domain the objects are specific RTs. An ontology's defined relationships among domain concepts also provide "slots" for named linkages between objects. As an example, the RT of finding someone's name on a device can be linked to the RT of searching for that name on LinkedIn.

When an ontology is populated with specific objects and explicit links are made between the objects, the ontology becomes a knowledge graph. Kurniawan et al. (p. 374, 2022) defined a knowledge graph as "…a directed, edge-labelled graph *G = (V, E)* where *V* is a set of vertices (nodes) and *E* is a set of edges (properties). A single graph is usually represented as a collection of triples *T = < s p o >* where *s* is a subject, *p* is a predicate, and *o* is an object." Each triple represents an object being linked to another object via a predicate. The predicate is also known as an object property.

Ontologies have been used in the cybersecurity domain for knowledge graph construction. Examples include the Situation and Threat Understanding by Correlating Contextual Observations (STUCCO) ontology (Iannacone et al., 2015) and the Unified Cyber Ontology (UCO) (Syed et al., 2016). Many open-source libraries also contain ontologies for the knowledge graph construction (Liu et al., 2022). Many types of graphs have been developed in the cybersecurity domain. As an example, Noel et al. (2015) developed a graph that predicts possible attack paths based on attack patterns and defense technologies. Ye et al. (2019) created a graph focused on attack paths to improve vulnerability assessments. There also have been many graphs developed that focus on security assessments (Zhang and Liu, 2020) and network security analytics (Noel, 2018).

Kurniawan et al. (2022) and Liu et al. (2022) identified a research need after examining reviews of previous cybersecurity knowledge-graph studies (Ding, Liu, and Zhu, 2020; Dong et al., 2020; and Yan and Liu, 2020). The identified research need consists of showing how knowledge graphs can be used to practical problems in cybersecurity at a broader scale than on a specific type of network with specific types of devices. Our research aimed to address this need by demonstrating a way to add to the Reconnaissance Tactic of ATT&CK. We first developed an ontology (called Recontology) and then filled the ontology with 102 RTs to form RT-Graph. Our research analyzed RT-Graph with statistical methods to draw insights about RTs. Using quantitative methods over graphs is considered to be critical for graph analyses (Noel et al., 2016).

3. Recontology and RT-Graph Development

3.1 Ontology Data Sources

Various sources can be used to create ontologies including industry-accepted frameworks, textbooks, standards, reports, domain expertise and others. For Recontology development, we used both ATT&CK, an industry-accepted framework in cybersecurity, and our domain expertise. Sugumaran and Storey (2002) considered domain expertise to be one of the most critical sources for ontology development. We hold over 10 years of experience in building cybersecurity ontology models and knowledge graphs in U.S. defense domain, 6 years of active-duty military experience, and multiple cybersecurity certifications including the Certified Information Systems Security Professional (CISSP) certification. We also have earned recognition for our reconnaissance work. Evidence includes being named on a university bug bounty program, having multiple Google Hacks accepted on the Google Hacking Database (GHDB), having submitted security reports to various university security teams, and having submitted a Common Vulnerabilities and Exposures (CVE) report to the National Vulnerability Database (NVD) which is pending review. In our reconnaissance work we have used various tools including Shodan, Nmap, Google, GHDB, Maltego, Metasploit and several more. We also use each tool in a variety of different ways.

3.2 Ontology Structure

Ontologies contain three main entities: classes, object properties, and data properties. Classes represent categories of objects. Object properties link together classes and data properties describe class characteristics. We used Protégé, an open-source ontology development tool, to create Recontology. Figure 1 shows Recontology classes. We created a super-class called Reconnaissance Tactic which contains all the RT classes. This Reconnaissance Tactic super-class maps directly to the ATT&CK Reconnaissance Tactic.

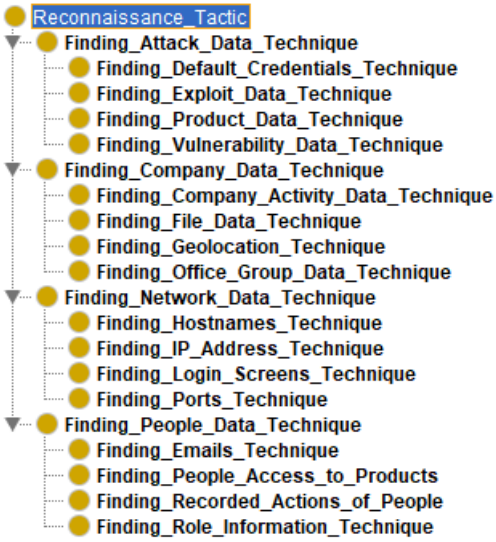


Figure 1: Classes in Recontology as Represented in Protégé

There are four upper classes that fall under the super-class Reconnaissance Tactic. The four upper classes are: Finding Attack Data Technique, Finding Company Data Technique, Finding Network Data Technique, and Finding People Data Technique. These four upper classes came from our years of experience in reconnaissance. Each upper-class contains sub-classes. There are a total of 16 sub-classes. Examples include Finding Default Credentials, Finding Hostnames, and Finding Emails techniques.

Recontology currently contains one object property called *providesDataFor*. We used this object property to show how results from one technique can provide data for another technique which is what we experience in the reconnaissance domain. This object property also directly relates to MITRE's (2022) conceptualization that RTs can "reveal opportunities for other forms of reconnaissance." In Recontology we explicitly model the connection among RTs.

Data properties provide an opportunity to name metadata about objects within classes. In Recontology, we use the following data properties: MITRE ATT&CK ID, Software, and Use. The MITRE ATT&CK ID data property is used to map Recontology RTs to MITRE ATT&CK technique ID numbers. The Software data property is used to define what tool is used for a specific Recontology RT. The MITRE (2022) definition of software is: "a generic term for custom or commercial code, operating system utilities, open-source software, or other tools…" The data property called Use provides a slot for stating the exact syntax to use for each RT.

Once objects were placed into the Recontology and were explicitly linked together via the object property *providesDataFor*, Recontology became RT-Graph. We used Protégé to enter the RTs and to explicitly link them together with direction (a RT points to another RT). When direction is modelled in a graph, the graph becomes a "directed graph." Across the 14 Recontology subclasses shown in Figure 1, there are 102 specific RTs within RT-Graph and the object property *providesDataFor* is specified over 200 times. We also filled out the three data properties (MITRE ATT&CK ID, Software, and Use) for each RT.

Figure 2 shows example RTs contained within the Finding Emails Technique class of RT-Graph.

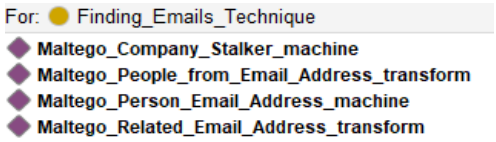


Figure 2: Objects Found in the Finding Emails Technique Class of RT-Graph

Figure 3 shows an example of the object property *providesDataFor* specified in RT-Graph. The example focuses on a specific Maltego RT.

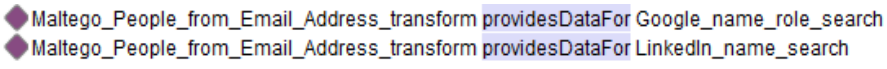


Figure 3: Example of the object property providesDataFor used in RT-Graph

After the ontology was developed in Protégé and the data for RTs (as well as object and data properties for each RT) were entered to form RT-Graph, we used the Protégé OWL2GraphML plug-in to transform the Web Ontology Language (OWL) file to a GraphML file. The OWL file format is a standard format used to save an ontology in Protégé and the GraphML file is an XML-based file that can be used by various graph-analysis programs such as Gephi.

4. Exploratory Data Analysis

We used exploratory data analysis (EDA) techniques to analyze the 102 RTs and the relationships among the RTs found in RT-Graph. With EDA, researchers attempt to find patterns and draw insights with data visual analytics, such as graph spatial-layouts, and with graph statistical algorithms (Langer and Meisen, 2019). Gephi, a graph visualization tool, provides various spatial layout algorithms and graph statistical algorithms for EDA.

4.1 Descriptive Metrics about the Graph

Initially, we focused on RT-Graph class and data property metrics. More specifically, we focused on the counts of RTs found in each of the 14 subclasses and the counts of software (such as Shodan, Nmap, Recon-ng and more) being used. We then examined trends across these metrics.

4.2 Spatial Algorithms for Visual Analysis for the Graph

For a directed graph like RT-Graph, there are various layout spatial-layout algorithms one can use within Gephi such as ForceAtlas, Fruchterman-Reingold, and Yifan Hu algorithms. Spatial layouts allow one to get a sense of groups of nodes (also known as subgraphs) and their characteristics, nodes that link subgraphs together, flow of information across graphs (as represented by arrows), and more. We chose the Force Atlas algorithm which is primarily used for directed graphs that have less than 10,000 nodes. Additionally, researchers can select filters to only show selected nodes based off statistical measures.

4.3 Statistical Data Analysis of Graph Features with Algorithms

Gephi provides several graph-statistical algorithms for data analyses. We explored the use of various algorithms in Gephi and selected the most appropriate ones for RT-Graph. The selected algorithms are named in Table 1 along with the measures they produce and meanings of the measures.

Table 1: Graph Statistical Algorithms

Algorithm Name	Algorithmic Measure	Explanation of Algorithmic Measure Results
Average Weighted Degree Algorithm	Weighted In-Degree	Weighs the number of incoming connections to a node; the higher the number, the more incoming connections
Average Weighted Degree Algorithm	Weighted Out-Degree	Weighs the number of outgoing connections; the higher the number the more outgoing connections
Average Weighted Degree Algorithm	Weighted Degree	Weighted total of incoming and outgoing connections; the higher the number the more "traffic" going through a node
Network Diameter Algorithm	Betweenness Centrality	The higher the number, the more critical the node is in the graph
Eigenvector Centrality Algorithm	Eigenvector Centrality	The higher the number, the more important a node is to the graph
Hyperlink-Induced Topic Search (HITS) Algorithm	Authority	The higher the number, the more valuable information the node holds
Hyperlink-Induced Topic Search (HITS) Algorithm	Hub	The higher the number, the more the node points to valuable, authoritative nodes

We analysed the results of each algorithm to look for trends within each algorithmic measure. We then compared the results across the algorithms to also look for trends. We also examined results across the descriptive graph metrics, spatial algorithms results, and statistical algorithms to draw insights and to examine patterns across the RT data.

5. Exploratory Data Analysis Results and Insights

5.1 RT-Graph Descriptive Metrics

RT-Graph consists of 16 subclasses that contain specific RTs. Examination of the RTs across the classes showed the following types of classes have the most RTs: finding vulnerability data (17), finding hostname data (14), finding IP address data (14), and finding port data (11). We also found there are a considerably smaller number of RTs for finding default credentials (2), log in screens (2), office-group data (1), and recorded actions of people techniques (1).

We examined the values of the data property "Software" for all 102 RTs, tracked the counts, and sorted the counts in Table 2.

Table 2: Counts of Tools Corresponding to RTs in RT-Graph

Shodan - 25	Device Access - 18	Nmap - 16	Recon-ng - 15
Google - 10	GHDB - 6	Maltego - 6	Metasploit - 2
Exploit DB - 1	LinkedIn - 1	NVD - 1	SpeedGuide - 1

We found that Shodan appeared the most (25) for all RTs. Upon examination of the Shodan RTs, we found that many of the techniques focus on finding a wide variety of data including screenshots, favicons, products, IP addresses, vulnerabilities, hostnames, ports, image, locations, and text in HTTP titles. We also found that a lot of the Shodan RTs rely on using hostnames, IP address or organizational names to find the aforementioned data. It was not surprising to use to see Shodan having the highest count among the tools. We use Shodan quite often in our reconnaissance work.

We also analyzed the access methods to the tools. Tools like Shodan, Device Access (meaning logging onto a device with credentials), Google, GHDB, ExploitDB, LinkedIn, the NVD, and SpeedGuide can all be used from a browser. However, some do require an account to be set up (Shodan, Device Access, and LinkedIn). We also found that 63 out of 102 tools (61.8%) can be accessed directly and freely from a browser while 38.2% require an install or download. Tools requiring an install include Nmap, Recon-ng, Maltego, and Metasploit. The above shows that our experience relies heavily on browser-based tools that are freely accessible. Also, all tools in RT-Graph are open source and free to use. Although these above results provide insights about the counts of data in RT-Graph, the descriptive results do not provide a clear picture about the relationships of and influence of RTs in RT-Graph.

5.2 Spatial Algorithm Results

Figure 4 shows the ForceAtlas spatial layout for RT-Graph. The parameter Weighted Out-Degree (described in Table 1) was selected to show the weighted directional-flow from RT to RT.

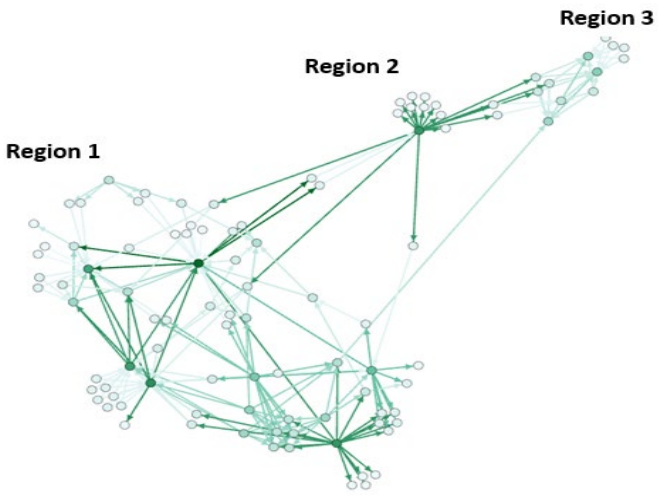


Figure 4: ForceAtlas Algorithm Spatial Layout of RT-Graph

The overall ForceAtlas spatial layout shows three large sub-graphs (or group of nodes). We labelled these as Regions 1, 2 and 3. In Region 1, many of these nodes are RTs that focus on finding things vulnerability, hostname, IP address and port data. This is the largest subgraph of RT-Graph. This made sense to us as the counts of RTs for finding these types of data are highest in RT-Graph as stated in Section 5.1

Region 2 represents both accessing a device and then searching the device for more data to use in other RTs. Finally, Region 3 represents searching for people data such as email addresses, access lists, and work roles. One can see also that some nodes sit in between Regions 1 and 2 that point /from various regions. While the spatial layout provided an overview of the whole RT-Graph, graph statistical-algorithms were needed to understand more details about RT-Graph.

5.3 Graph Data Analysis Results and Interpretation

Table 3 shows the results for Weighted In-Degree, Weighted Out-Degree, and Weighted Degree measures calculated by the Average Weighted Degree algorithm.

Table 3: Weighted In-Degree, Out-Degree and Degree Results for RT-Graph

Algorithmic Measure (Maximum Value for RT-Graph)	Top Results From RT-Graph (Measure Value)
Weighted In-Degree (Maximum = 20.0)	• Google search for product name (20.0) • ExploitDB search for exploits (17.0) • Nmap Intense scan for all TCP ports (15.0)
Weighted Out-Degree (Maximum = 19.0)	• Nmap Quick Scan for hostname (19.0) • Device access with credentials (17.0) • Shodan search with organization name to find hostname (11.0) • Nmap Quick Scan for IP address (10.0)
Weighted Degree (Maximum =25)	• Google search for product Name (25) • Nmap Intense scan for all TCP ports (20) • Device access with credentials (19) • Nmap Quick Scan for hostname (19)

For the Weighted In-Degree measure, the top three nodes are searching for product, exploit data, and the Nmap Intense scan. This made sense to us because of key information that is needed to conduct these RTs: product name for product search, CVE or product info for the exploit search and hostname or IP address for the Nmap search. Many other RTs can provide these input data. For the Weighted Out-Degree, we saw that Nmap Quick Scans, Device Access, and Shodan search provide data that are used by a lot of other RTs. The Weighted Degree showed that the product name search, Nmap scans, and device access have a lot of "traffic" going through with "traffic" representing the total and in and out links. We use these RTs often in our reconnaissance work.

Table 4 shows values of the top-ranked RTs based on the Betweenness Centrality, Eigenvector Centrality, Authority, and Hub statistical results.

Table 4: Betweenness Centrality, Eigenvector Centrality, Authority and Hub Results for RT-Graph

Algorithmic Measure (Maximum Value for RT-Graph)	Top Results From RT-Graph (Measure Value)
Betweenness Centrality (Maximum = 1824.8)	• Google search for product name (1824.8) • Device access with credentials (1673.0) • Google search for default credentials (816.0) • Google search for technical manual with default credentials (816.0)
Eigenvector Centrality (Maximum = 1.0)	• ExploitDB search for exploits (1.0) • Google search for product (0.58) • NVD search (0.36) • Metasploit product search (0.36)
Authority (Maximum = 0.34)	• Nmap Intense scan for all TCP ports (0.34) • Nmap Intense scan with UDP ports (0.33) • GHDB for error messages, files with usernames, footholds, sensitive directories, vulnerable files, and vulnerable servers (all same value of 0.32)
Hub (Maximum = 0.50)	• Nmap Quick Scan for hostname (0.50) • Recon-ng Brute Hosts module for hostnames (0.42) • Recon-ng HackerTarget module for hostnames (0.40) • Shodan Search with organization name to see hostnames (0.35) • Google Search to find hostname (0.32)

The Betweenness Centrality measure shows the nodes in a graph that play a key role in linking parts of the graph together. The Google Search for Product Name RT and the Google Searches for Credential RTs (two of them) play a key role in linking Region 1 with Regions 2 and 3. Also, when a device is accessed it can be used to find a lot of other data such as access lists, logs, and more.

The Eigenvector Centrality measure shows the most critical nodes in a graph. ExploitDB holds a considerably higher value that other RTs. This made sense to us as the goal of reconnaissance is to support exploitation (MITRE, 2022). Additionally, NVD search and Metasploit product search also have higher values than most other nodes. These two also contribute considerably to supporting exploitation. And to use these two RTs, one needs to know the product name which is why the Google Search for Product RT is ranked high for Eigenvalue Centrality. We also saw that the that top four results are all for different tools and that the top three are all browser-based tools.

The Authority measure estimates the value of the content of the node. Nmap intense scans ranked at the top. They provide key data such as operating system (OS) version, service version, ports status and more. We see that six GHDB techniques also rank at the top. These techniques provide critical inside information across an organization beyond one device. It was interesting to see only two tools at the top of this measure and that 6 of the top 8 can all be used over the browser.

Finally, the Hub measure estimates the value of the links that nodes send to other nodes. It was interesting to see four different tools in the top five results. And it was interesting to see that each of these top-five RTs focused on finding hostnames. That shows how critical finding hostnames is in our reconnaissance body of knowledge.

5.4 Overall Comparison across Descriptive Metrics and Graph Analysis Algorithms

As we examined the results across the various data analysis techniques, we noted various trends. Overall, the RT of Google Search for Product Name represents a significant RT in our body of RT knowledge. It ranked 1 in the Betweenness Centrality, Weighted In-Degree, and Weighted Degree measures. It also ranked 2 in the Eigenvector measure. This shows how critical researching a product is in our reconnaissance approach.

After examining the top Eigenvector values in more detail, we found that 9/10 (90%) of the techniques rely on tools that can be used directly in the browser. Metasploit was the only one that could not be used directly in the browser. There are other ways to search for modules or exploits. however, our team uses Metasploit within Kali Linux.

We also see that Nmap is a powerful tool in our body of reconnaissance knowledge. Nmap as a tool that was associated with RTs that were ranked 1 for Weighed-In Degree, the Hub measure, and Authority. Also, Nmap was associated with the rank 2 and 4 for Weighed-In and Authority respectively. This made sense to us since we often use Nmap to find detailed information in reconnaissance.

We also looked at the graph data analysis results side-by-side with the descriptive metric results. It was interesting to note that ExploitDB is only used for one RT. However, this one RT became the most critical node in RT-Graph according to the Eigenvector Centrality metric. It was also interesting to know that GHDB is associated with six RTs and all six (100 percent) appeared as top-ranked Authority RTs. Finally, Metasploit is associated with two RTs but ranked fourth in the Eigenvector centrality metric.

One unexpected result was encountered. We often use Shodan in various RTs and value the information that we find from it. However, Shodan was not associated with many critical nodes based on graph statistical-algorithm results. We realized though we have used Shodan for 25 different RTs. We reasoned that perhaps because we use many Shodan RTs, it is difficult for one RT to stand out.

5.5 Future Work and Applications

The results of this research demonstrated which RTs our team relies on the most. It prompted us to explore other RTs and other reconnaissance tools such as Amass and SpiderFoot. Also, the research encouraged us to examine which RTs under ATT&CK are missing from our body of knowledge. Example ATT&CK RTs that we plan on exploring include Phishing for Information, Searching Closed Sources, and Searching Social Media. Furthermore, we plan on examining MITRE ATT&CK Tactics beyond Reconnaissance since Reconnaissance is only one of 14 tactics. Example tactics include Execution and Exfiltration. We also have formed a method for documenting new RTs and ingesting the new RT data into RT-Graph. We will use a shared Excel spreadsheet to track new tools, techniques, and syntaxes used. Then we will use the Protégé plugin named Cellfie to ingest the

Excel data into RT-Graph. Finally, as new knowledge is added to RT-Graph, our current RT-Graph can be compared to a future instance of our RT-Graph. Then the two RT-Graphs can be quantitatively and qualitatively compared to examine knowledge changes.

We see potential real-world applications of RT-Graph as our research continues to mature. For example, each individual team member brought different expertise across various RTs. Informally, various team members tested out new RTs that there weren't familiar with to expand their knowledge. In the future, we could examine how graph development or analyses increases one's knowledge. Also, we can search RT-graph by tool or by RT. The results can of the searches can serve as a checklist for completing reconnaissance projects for both experienced and new team members especially since each RT also comes with specific syntax to use.

6. Conclusion

ATT&CK provides a model of adversarial tactics and techniques including the Reconnaissance Tactic and its associated reconnaissance techniques (RTs). ATT&CK conceptualizes that RTs can inform other RTs. However, ATT&CK does not explicitly link RTs together. We created an ontology (Recontology) to model the Reconnaissance domain and then turned the ontology into a knowledge graph (RT-Graph) to explicitly link together 102 RTs. RT-Graph represents our body of knowledge in the reconnaissance domain. Using descriptive statistics, a graph spatial-layout algorithm, and graph statistical algorithms, we found the RTs that are most critical in RT-Graph, that have the most traffic going through them, that link main areas of RT-Graph together, that hold the most authoritative information, and the point to the most authoritative sources. Our research provides an approach for explicitly linking RTs together for using exploratory data analysis techniques to examine the RT knowledge graph. Future research will focus on building RT-Graph and comparing future instances of RT-Graph to the current one in order to examine knowledge growth.

Acknowledgements

The researchers acknowledge Addison Lilholt, from the Baldwin School, and Jeremy Slaven, from Drexel University, for their valuable feedback on our knowledge graph.

References

Aviad, A., Wecel, K. and Abramowicz, W. (2015) "The Semantic Approach to Cyber Security towards Ontology Based Body of Knowledge", *European Conference on Cyber Warfare and Security*.

Ding Z., Liu K., Liu B., and Zhu, X. (2020) "Survey of Cyber Security Knowledge Graph*", Journal of Huazhong University of Science and Technology,* Vol 49, No. 7, pp 79-91.

Dong, C., Jiang, B., Lu, Z., Liu, B., Li, N. and Ma, P. (2020) "Knowledge Graph for Cyberspace Security Intelligence: A Survey", *Journal of Cyber Security* Vol 5, pp 56–76.

Georgiadou, A., Mouzakitis, S., and Askounis, D. (2021) "Assessing MITRE ATT&CK Risk Using a Cyber-Security Culture Framework", *Sensors*, Vol 21, No. 9, pp 1-14.

Gruber, T.R. (1993) "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, Vol 5, No. 2, pp 199-220.

Iannacone, M., Bohn, S., Nakamura, G., Gerth, J., Huffer, K., Bridges, R., Ferragut, E. and Goodall, J., "Developing an Ontology for Cyber Security Knowledge Graphs", *The 10th Annual Cyber and Information Security Research Conference*, pp 1-4.

Kurniawan, K., Ekelhart, A., Kiesling, E., Quirchmayr, G., and Tjoa, A. M. (2022). "KRYSTAL: Knowledge Graph-Based Framework for Tactical Attack Discovery in Audit Data" *Computers & Security*, Vol 121.

Langer, T., and Meisen, T. (2019) "Towards Utilizing Domain Expertise for Exploratory Data Analysis", *The 12th International Symposium on Visual Information Communication and Interaction,* pp 1-5.

Liu, K., Wang, F., Ding, Z., Liang, S., Yu, Z., and Zhou, Y. (2022) "Recent Progress of Using Knowledge Graph for Cybersecurity", *Electronics*, Vol 11*,* No. 15.

MITRE (2022) "Reconnaissance Tactic-TA0043", viewed January 1, 2023 <https://attack.mitre.org/tactics/TA0043/>

Noel, S., Harley, E., Tam, K.H., Limiero, M., and Share, M. (2016) "CyGraph: Graph-Based Analytics and Visualization for Cybersecurity." In *Handbook of Statistics*, Vol 35, pp 117–167.

Noel S. (2018) "A Review of Graph Approaches to Network Security Analytics", In *From Database to Cyber Security*, pp 300-323.

Noel, S., Harley, E., Tam, K. H., and Gyor, G. (2015) "Big-Data Architecture for Cyber Attack Graphs: Representing Security Relationships in NoSQL Graph Databases", *IEEE Symposium on Technologies for Homeland Security*.

Steinberg, A. N. (2017) "Foundations of Situation and Threat assessment", In *Handbook of Multisensor Data Fusion,* pp 457-522.

Sugumaran, V. and Storey, V. C. (2002) "An Ontology-Based Framework for Generating and Improving Database Design" *International Conference on Application of Natural Language to Information Systems,* pp. 1-12.

Syed Z, Padia A, Finin T, Matthews, L. and Joshi, A. (2016) "UCO: A Unified Cybersecurity Ontology"*, 30th AAAI conference on Artificial Intelligence*.

Yan Z. and Liu J. (2020) "A Review on Application of Knowledge Graph in Cybersecurity", *International Signal Processing, Communications and Engineering Management Conference (ISPCEM*), pp.240-243.

Ye, Z.W., Guo, Y.B., Li, T. and Ju, A.K. (2019) "Extended Attack Graph Generation Method Based on Knowledge Graph", *Computer Science*, Vol 46, No. 12, pp 165-173.

Zhang, K. and Liu, J. (2020) "Review on the Application of Knowledge Graph in Cyber Security Assessment" *IOP Conference Series: Materials Science and Engineering*, Vol. 768.

Zero Trust is Not Enough: Mitigating Data Repository Breaches

J.S. Hurley
ODNI, Bethesda, USA

john.s.hurley@odni.gov

Abstract: Successful mission operations depend on the ability of an organization to collect, manage, analyze, and secure its data. Traditional network frameworks have become less appealing because they rely on a "trust but verify" paradigm that does not stand up well against the advanced tools and techniques of modern cyber attackers. The Zero Trust Framework has emerged as a logical replacement but unfortunately does not adequately address the trustworthiness of data in data repositories. This broader view is important because data repositories have become arguably, the most prominent means for data sharing across many sectors around the globe. Unfortunately, data repositories are also undergoing widespread malware attacks and, in some cases, data critical to national security can be impacted. In this study, we propose a potential framework that relies more on data lineage, end-to-end metadata, and the use of machine learning tools and techniques to reduce and possibly mitigate the problems with data repositories that Zero Trust Frameworks fail to address.

Key Terms: Zero Trust, Data Repository, Frameworks, Machine Learning.

1. Introduction

Senior leaders must make decisions that can immensely impact operations and Mission success. As might be expected, decisions and how they are implemented can have significant implications to Mission success. Hence, the accuracy, accessibility, and availability of the data are crucial in the decision-making process. Data scientists facilitate an improved and informed decision-making process through the unique additional insight they provide. A data repository is an important element of a data scientist's toolbox wherein select datasets can be placed at an acknowledged centralized, isolated location, made available to users to validate existing and also generate new results. Data repositories have the additional benefit of serving as the primary mechanism in which data scientists facilitate collaboration and information sharing between users.

Application stacks are a dominant way that organizations carry out many of their key business functions. Third-party apps have become a staple of how we interact with data and services in both cloud and mobile-driven environments. There is general consensus that third party apps have immense benefits. However, forethought must also be given to the potential risks of third-party apps (Wang, et al., 2020).

Data repositories serve as the central point in which large amounts of computer code are stored and managed. As a result, security of the repository is a primary requirement and concern. However, herein lies a major inherent conflict. Although security is a necessary requirement, too many restrictions can be counterproductive because the key to collaborative and open-source development depends on the accessibility to data repositories. Recent data repository breaches seem to be signaling a growing and troubling trend. According to Sonatype, the volume of malicious attacks on open-source code repositories over the last three years has hit triple digit growth (Scott & Spitse, 2022)! It is reported that an almost 700% rise in attacks designed to plant malware in software components was detected. Almost every business relies in some way or another on access to open-source data. Compromised or lost credentials or breach of the underlying service could allow attackers to modify a code base without permission. Attacks can be subtle or disguised, so it is important to review any commits, changes or pull requests to understand security implications. The manner in which credentials are injected, managed, and stored into the computing environment is also critical. For example, the Uber hacks were all done by stealing server passwords that Uber had accidentally posted on GitHub repositories (twice!) (Fair, 2018).

Most of the recently recorded attacks on data repositories have largely been accomplished without exposing the passwords of users. Attackers have gained access through "access" tokens that authorize the sharing of specific user account information (Curwin, et al., 2023). The latest attacks on data repositories seem to have a common thread, i.e., the attacks seem to be coming primarily through third party access provided by OAuth, the open standard for token-based authentication and authorization. Through OAuth request links, recipients can be tricked into illicit grants, e.g., consent phishing emails that can enable access to attackers via API resources. In this case, targeted users unknowingly grant permissions that allow attackers to make API calls on their behalf through attacker-controlled apps. Unfortunately, access to cloud SaaS environments is obtained with relative ease due to end user-granted permissions occurring without much scrutiny. Permissions can be granted by end

users simply via a permissions request submitted from the third-party app. Similar problems can also occur through browser extensions via application performance interfaces (APIs).

It is important to note that new vulnerabilities can be introduced any time a change is made to a codebase. As a result, it is prudent to check for net vulnerabilities before merging a pull request. However, even this can be challenging in containerized applications where vulnerabilities can enter any container layer from the operating system to the application. Pre-existing container and dependency vulnerabilities in the target branch must be known in order to identify new vulnerabilities in a pull request. A comparison between the pull request and target branch images can provide a view into new vulnerabilities that might be introduced in containers and dependencies (Kaur, et al., 2021). Additionally, compromises of packages or libraries can lead to downstream infections. *Of particular concern is that containers and the microservices they provide have increased the surface area almost exponentially. Container image vulnerabilities, unauthorized access, and weak firewalls are a few of the challenges that containers face* (White, 2022).

Zero Trust is a security framework that continuously authorizes, authenticates, and validates users in order to be granted and maintain access to data and applications (Raina, 2022). Zero Trust assumes that networks can be local, cloud-based, or some combination, i.e., there is no traditional network edge. NIST Special Publication (SP) 800-207 is currently the de facto standard for Zero Trust Architectures for both industry and government agencies (Rose, et al., 2020). *Without question, Zero Trust provides significant benefits and performs admirably, in protecting against many cyberattacks, reducing attack surfaces, and simplifying infrastructure requirements. In addition, Zero Trust Architectures (ZTAs) are designed to better meet many new business requirements driven by digital transformation that can increase risk. This benefit of zero trust belies the fact that traditional perimeters are complex and can increase risk. Zero Trust, however, does have its challenges, e.g., Zero Trust must be slowly phased into infrastructure because it is much more restrictive and can require an entirely different business model than what might already be in place* (Chimakurthi, 2022).

In this study, the focus will be on the design of a potential framework that relies on the use of machine learning tools and techniques to reduce and possibly mitigate some of the problems with current and future breaches that can plague data repositories. First, the study begins with a discussion of data repositories and the value they have in the decision-making process. Next, the concept of Zero Trust and how it significantly improves a number of past issues with regards to security are considered. A series of recent breaches will then be discussed in detail while sharing suspected ways that attackers gained unauthorized entry into some data repositories. The Results and Discussion session will then follow with a proposed framework that utilizes various data science techniques to address the possibility of mitigating the breaches in data repositories. Specifically, the focus will be on how data science technologies can be used to determine if OAuth apps are either “risky” or “allowed”. The issues of commits and pull requests (PRs) are important but will be considered only briefly in this work. In those discussions, the context will be on how pre-existing container and dependency vulnerabilities in the target branch can be passed on through pull requests and commits. More extensive investigations will be the subject of future work.

2. Data Repositories

Data repositories represent a large database infrastructure in which datasets are collected, managed, and stored for analysis, reporting, and sharing. Data repositories provide immense value due to their ability to store, make accessible, organize, and archive datasets in a logical manner. From a scientific viewpoint, data repositories are critical because they enable data sharing, one of the primary mechanisms for advancing knowledge in a timely manner. Data repositories have added benefits, including serving as the centralized location that makes it possible to pull in data from multiple sources and improve the quality of preserved and aggregated data. Some of the most prominent repositories are placed on open-source end-to-end software development platforms such as GitHub, Bitbucket, and GitLab. As data grows, the database management system must be scalable to accommodate the inherent slowing down of the system. Users typically interact with code repositories through a web application or command-line utilities such as Git.

3. Zero Trust

Zero Trust is a strategic approach that eliminates implicit trust while continuously validating most stages of a digital interaction. Zero Trust is designed to enable digital transformation and protect modern environments by leveraging network segmentation; preventing lateral movement; providing Layer 7 threat prevention; simplifying granular “least access” policies; and using strong authentication methods (Souppaya, et al., 2022).

Network complexity, overlays and dynamic IPs in combination with the limitations of traditional firewalls can cause problems with traditional security approaches (Fearn, 2020). Zero Trust Architectures (*ZTAs), however, fall short in dealing with some issues. In particular, one of the biggest challenges of zero trust is that it can be very complex to implement, given that every application, device, and user must be authenticated and authorized. For organizations with large numbers of users, this requirement can be especially challenging because of the layer of complexity added. Next, a discussion of some of the recent data repository attacks and resulting challenges are considered.*

4. Attacks

Credentials used or misused in the network account for more than 80% of all cyber attacks. Attention to greater account integrity; adherence to organizational rules; avoidance of high-risk shadow IT services; password security; integrity of accounts; adherence to organizational rules; and avoidance of high-risk shadow IT services should all, at a minimum be actuated (Raina, 2022). On August 3rd, software developer Stephen Lacy discovered a data repository attack involving bad actor cloning and adding malicious code to more than 35,000 GitHub repositories, while still keeping the code's original source code. Almost 40% of the infected repositories originated from a single organization, referred to as a "redhat-operator-ecosystem" on the site. Also, the code could download additional malware from a third-party site that allowed further exploitation of any environment or application that was using the malicious cloned code originally introduced to the GitHub repositories. The weaponized code could potentially lead developers to accidentally download cloned code repositories which contained malicious code (Powell, 2022).

In 2022, the identification and authentication giant, Okta, reported that it was informed by GitHub of "suspicious" activity in its code repositories. Okta concluded that hackers gained unauthorized access to copy code repositories. The breach incidents reinforced how crucial logs are to be put in place and continuously monitored to detect if any of Okta's systems were inappropriately accessed and if any data had been exfiltrated (Page, 2022). Undiscovered flaws and vulnerabilities on open-source GitHub repositories are increasingly becoming a major concern. Delayed or inconsistent code commits along with improper scanning also create risk as repositories age. The identity and access management platform, Okta, reported on December 2022 that its source code repositories had been inappropriately accessed and copied. Similar breaches and unauthorized access to code bases has also been reported by password manager, LastPass, in which a backup of its customer vault data was copied, impacting over 33 M registered users (Kapko, 2023).

5. Results and Discussion

It is important to reiterate that two primary mechanisms that can contribute to breaches on data repositories are the subject of discussion in this study. Third-party access (via the authorization and authentication standard, OAuth) and when changes, commits, or pull requests are made to a codebase. It is a main premise of this paper that a new framework that incorporates the impact of both possibilities must be introduced and implemented to reduce and possibly mitigate the challenges of data repository breaches.

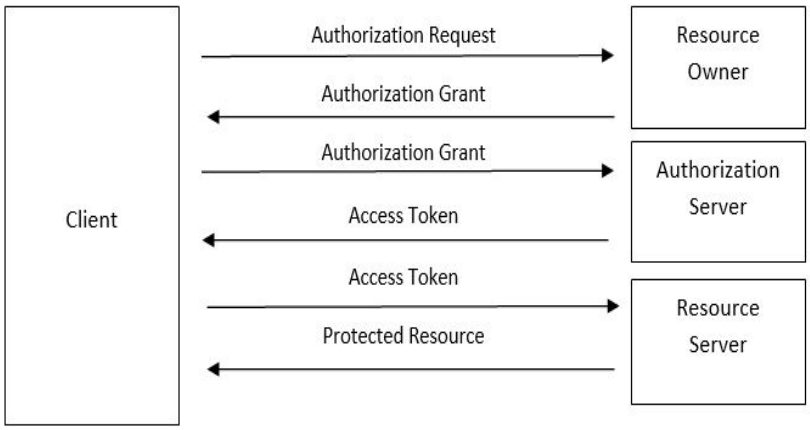


Figure 1: Access Token Issuance Process, (Shah, 2022)

In Figure 1., the process in which access tokens are generated and passed on to the third-party apps is shown. In this process, we must pay attention to possible malicious and "risky" third-party apps that could be introduced and how they might be detected and mitigated. It is important for users to review processes and procedures within their environments. Hence, filters should be set to Permission levels of high severity and uncommon community apps that are potentially very risky should not be allowed in the system. Under Permissions, all options should be selected that are particularly risky in a specific context. In particular, email access permissions should be carefully reviewed, especially when full access to all mailboxes is allowed (Curwin, et al., 2023). In Table 1., some key tests to determine if OAuth apps are risky or suspicious, or be allowed, or require further testing are shown below.

Table 1: OAuth Testing for Risk and Suspicious Apps

Testing (Phase 1)	Risky	Allowed
Email-related access users that have unnecessary permissions	x	
Apps authorized by external users that do not fit within organization's security standards and policies	x	
Apps authorized by a small number of users	x	
Apps with suspicious name, publisher, or URL	x	
Apps and Targeted apps that do not have current last authorization date	x	
Common or Often Used		x
Only Permissions linked to app's purpose allowed		x
Requires high privileges or administration consent	x	
Activities performed by the app are appropriate		x
Users consent to an app		x
If the OAuth 3rd party app satisfies the above allowed criteria, then they can be allowed. Otherwise, a second phase of testing should be implemented. In particular, an examination of the app's name, publisher, and URL in different app stores should be carried out with a focus on including:		
Testing (Phase 2)		
Low number of downloads	x	
Low rating or score or bad comments	x	
Suspicious publisher or URL	x	
Date of last update not recent	x	
Irrelevant permissions	x	
OAuth apps that fail to pass the second phase of testing should be disqualified and considered too risky		

If the app is determined to be risky, either manual or automatic remediation methods can be used to revoke the app or access privileges of the specific user of the app.

In the recently reported breaches, infections from third party apps were a primary contributor to the breaches. Immediate attention must be given because organizations are increasingly collecting, storing, and using data to gain critical insight to make the most effective business decisions. Data repositories have been viewed as a mechanism that could overcome many of the issues noted given their access and long-term storage capabilities that make them well-suited as a sustainable information infrastructure. In this study separate applications will be placed into clusters that can be categorized as "allowed" or "risky" before they are incorporated into the data repositories to reduce the likelihood of vulnerabilities.

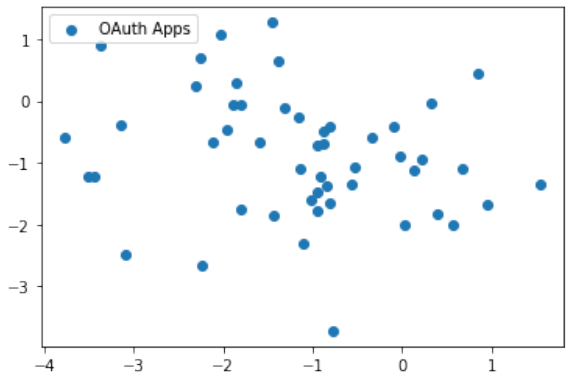


Figure 2: Sample Cluster 100 points of OAuth Apps vs Allowed Apps

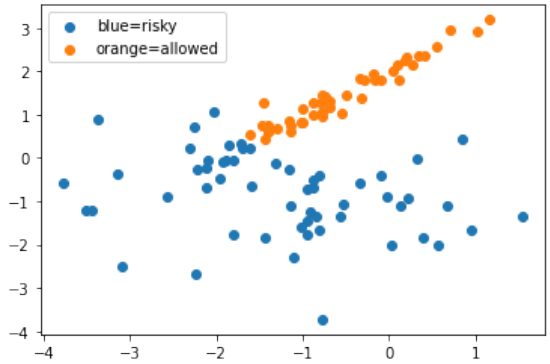


Figure 2. Sample Cluster 100 Points of OAuth Apps

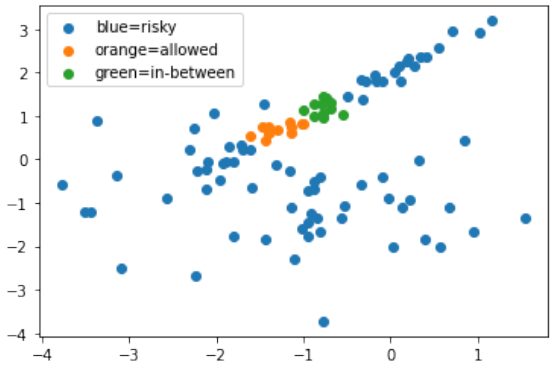


Figure 3. Agglomerative Clustering Applied for Risky vs Allowed Apps

Figure 4. DBScan Clustering Applied for Risky vs Allowed Apps

In Figure 2., there are 100 OAuth apps (points) randomly selected. In Figure 3., the points are categorized as either "risky" or "allowed" according to the criteria listed in Table 1. In Figure 3., the apps considered "risky" are colored blue (note: the blue points in Figure 1.are just arbitrary points not yet categorized) and those that are allowed are colored "orange". In this study, the Agglomerative clustering algorithm (ACA) was used based on the features provided in Table 1. to categorize the apps because it bases group objects in clusters in terms of their similarity. This algorithm is considered a greedy algorithm because of how it determines the appropriate mergers and splits. In Figure 2., the "risky" vs "allowed" apps are identified and separated to enable users to determine which apps should be accepted into the repositories. The ability to make this determination beforehand is considered critical to maintaining the desired security posture of the data repositories so that users and owners can feel confident that vulnerabilities are not being passed on. It is usually prudent to consider at least another clustering algorithm for comparison of the results. In this study, the density-based (DBscan) algorithm was selected for comparison vs the ACA algorithm. The DBScan representation is interesting because it identifies an "intermediate" class between the risky and allowed points. The ACA was chosen, to get a clear delineation between "risky" and "allowed" OAuth apps.

In addition to the OAuth risky third-party apps, the other challenge noted to impact data repository security, i.e., the way that pull requests (PRs), commits and other changes are allowed within data repositories (Yu, et al., 2018). Researchers have recently started to investigate the pull request workflow because the evaluation of PRs is an iterative, complex process involving many stakeholders (Gousios, et al., 2014), (Tsay & Dabbish, 2014), (Tsay, et al., 2014), (Tsay, et al., 2014). Unfortunately, the factors that can influence PR acceptance (decision to merge) and latency (evaluation time) are still not well understood. In addition, PR evaluations have not yet been fully examined in terms of continuous integration (CI) (Jiang, et al., 2017).

In Figure 3, shown below, a GitHub pull-request operation flow diagram is provided.

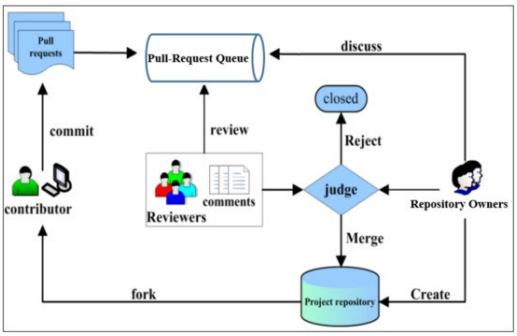


Figure 3. Overview of Pull-Request Mechanism in GitHub, (Salman, et al., 2022)

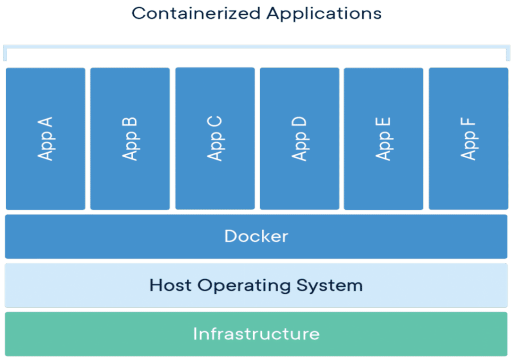


Figure 4. Docker Containerized Applications (See What is a Container? | Docker)

The continuous and parallel distributed development make managing changes in software development very difficult (Nerur, et al., 2005) (Dabbish, et al., 2012).

Containers enable code to be rolled out more quickly and can enable app development time to be shortened while using fewer storage and compute resources. Containers are usually accompanied by an orchestration tool, which essentially controls the container. Most of the orchestrator and container vulnerabilities correspond to the classic vulnerabilities usually found in applications and operating systems, including: access and authorization, API server access, image vulnerabilities, inter-container network traffic, and other (Mell & Pariseau, 2023).

What is needed is a framework that makes it possible to verify the authenticity of an artifact (in this case a container image) and check for anomalous behavior (vulnerabilities) due to cyber attacks. The image contains code, runtime, system tools, system libraries and settings as shown in Figure 4. In this work, given the interest is in determining if a container image has a vulnerability, it is proposed that natural language processing (NLP) be applied to container image recognition to identify vulnerabilities. In this approach a standard "WordPiece" tokenizer that tracks fractions of words would be used and converted into select features of the image. The WordPiece approach could cluster traffic in an analytically meaningful way. The benefit is that no preprocessing would be required because raw input could be tokenized directly. Additionally, a key takeaway is that the proposed model would group the artifacts by behavior and automatically learn to characterize "normal" vs. altered behavior. In this approach, the system would learn enterprise normality. This approach also uses embeddings to identify adversarial bias in which a key takeaway might be the ability to tell the degree of threat just by its relative position in embedding space. Words generally arrange themselves spatially along a continuum of meaning. Embeddings can be used to answer complex what-if questions that are very hard to do in some other cases. What-if Analysis is employed using Embedding Vector Math. Connection embeddings show adversarial bias, wherein the difference between embeddings of two communicating containers is an embedding for their connection. In most cases, most of the suspicious traffic forms a cluster. The benefit is that the model can separate traffic into analytically meaningful types without supervision (i.e., no labels).

6. Conclusion

In this study, a couple of the factors have been reinforced as key contributors to the recent data repository breaches, including: the risk associated with suspicious OAuth apps and the challenging process of commits and pull requests. Most of the attention in this study was placed on the OAuth apps. In particular, there are a number of criteria used to determine when an app should be allowed or considered too risky. To date, there is very little effort to isolate and define the "bad" OAuth apps. A table was presented that does the isolation and specifically defines parameters that must be addressed before apps can be allowed into data repositories. In the study, 100 arbitrarily selected apps were successfully clustered based on pre-defined criteria using Agglomerative clustering and DBscan algorithms to define specifically those apps that are allowed and those that should be excluded. Zero trust architectures, without a doubt have improved the security posture of most current computer environments. However, some challenges still remain in terms of when it will be ready to address data repository vulnerabilities. A new approach introduced by PNNL to visualize normal enterprise behavior via

natural language processing applied to address the huge data in cyber logs because cyber log data is different and there are domain specifics that can help or hinder analysis. The approach is considered in this work to address the vulnerabilities in containers, OSs, and dependencies (Ranade, et al., 2021) and mitigate vulnerabilities passed on from container to that can get pulled into the host computer. Future efforts will focus on the commits and pull requests challenges and the development of a machine learning pipeline to include both the risky Auth apps and commits/pull request challenges. In addition, other machine learning algorithms will be considered to compare performance with the decision tree algorithm.

Bibliography

Chimakurthi, V., 2022. The Challenge of Achieving Zero Trust Remote Access in Multi-Cloud Environment. *ABC Journal of Advanced Research,* 9(2), pp. 1-13.

Curwin, D. et al., 2023. *Investigate and Remediate Risky OAuth Apps.* [Online]
Available at: https://learn.microsoft.com/en-us/defender-cloud-apps/investigate-risky-oauth
[Accessed 19 01 2023].

Dabbish, L., Stuart, C., Tsay, J. & Herbsleb, J., 2012. Social Coding in GitHub: Transparency and Collaboration i an Open Software Repository. *Association for Computing Machinery*, 12-17 02, pp. 1277-1286.

Fair, L., 2018. *FTC Addresses Uber's Undisclosed Data Breach in New Proposed Order.* [Online]
Available at: https://www.ftc.gov/business-guidance/blog/2018/04/ftc-addresses-ubers-undisclosed-data-breach-new-proposed-order
[Accessed 19 01 2023].

Fearn, N., 2020. *How to Apply Zero-Trust Models to Container Security.* [Online]
Available at: https://www.computerweekly.com/feature/How-to-apply-zero-trust-models-to-container-security
[Accessed 07 02 2023].

Gousios, G., Pinzger, M. & and v.Deursen, A., 2014. An Exploratory Study of the Pull Based Software Development Model. *International Conference on Software Engineering, ICSE*, pp. 345-355.

Jiang, J. et al., 2017. Analyzing Attributes for More Accurate Commenter Recommendation in Pull-Based Development. *Science Direct,* 84(Information and Software Technology), pp. 48-62.

Kapko, M., 2023. *Open-source Repository Risk Amplified on GitHub.* [Online]
Available at: https://www.cybersecuritydive.com/news/open-source-repository-risk/640333/
[Accessed 19 01 2023].

Kaur, B., Dugre', M., Hanna, A. & Glatard, T., 2021. An analysis of security vulnerabilities in container images for scientific data analysis. *Gigascience,* 10(6), pp. 1-7.

Nerur, S., Mahapatra, R. & Mangalaraj, G., 2005. Challenges of Migating to Agile Methodologies. *Communications ACM ,* Volume 48, pp. 72-78.

Page, C., 2022. *Okta Confirms Another Breach After Hackers Steal Source Code,* San Francisco: TechCrunch.

Powell, O., 2022. *GitHub Supply Chain Attack Could Affect 83 Million Developers.* [Online]
Available at: https://www.cshub.com/attacks/news/github-supply-chain-attack-could-affect-83-million-developers
[Accessed 12 01 2023].

Raina, K., 2022. *www.crowdstrike.com.* [Online]
Available at: www.crowdstrike.com/cybersecurity-101/zero-trust-security/
[Accessed 12 01 2023].

Rose, S., Borchert, O., Mitchell, S. & and Connelly, S., 2020. *Zero Trust Architecture.* [Online]
Available at: https://www.nist.gov/publications/zero-trust-architecture
[Accessed 12 01 2023].

Salman, H., Alshar, Z. & Seriai, A., 2022. *Automatic Identification of Similar Pull-Requests in GitHub's Repositories Using Machine Learning.* [Online]
Available at: https://www.mdpi.com/2078-2489/13/2/73
[Accessed 19 01 2023].

Scott, T. & Spitse, N., 2022. *Third Party Risk is Becoming a First Party Challenge.* [Online]
Available at: dhttps://www2.deloitte.com/ca/en/pages/risk/articles/reduce-your-third-party-risk.html
[Accessed 05 02 2023].

Shah, S., 2022. *How OAuth 2.0 Supports Third-Party Applications to Authorize HTTP Service?.* [Online]
Available at: https://www.einfochips.com/blog/the-use-of-oauth-2-0-framework-for-third-party-applications-to-authorize-http-service/
[Accessed 19 01 2023].

Souppaya, A., Scarfone, K., Symington, S. & Barker, W., 2022. *Implementing a Zero Trust Architecture,* Gaithersburg: NIST.

Tolbert, J., 2020. *A Look at NIST's Zero Trust Architecture.* [Online]
Available at: https://www.kuppingercole.com/blog/tolbert/a-look-at-nists-zero-trust-architecture
[Accessed 12 01 2023].

Tsay, J., Dabbish, L. & Herebsleb, J., 2014. *Let's Talk About It: Evaluating Contributions through Discussions in GitHub.* Kansas City, ACM.

Tsay, J. & Dabbish, L. H. J., 2014. *Infleuence of Social and Technical Factors for Evaluating Contribution in GitHub.* Kansas City, ACM.

Wang, Y. et al., 2020. *An Empirical Study of Usages, Updates and Risks of Third-Party Libraries in Java Projects.* Cyprus, IEEE International Conference on Software Maintenance (ICSM).

White, D., 2022. *How Zero Trust Models Work in Container Security.* [Online] Available at: https://thenewstack.io/how-zero-trust-models-work-in-container-security/ [Accessed 07 02 2023].

Yu, S. et al., 2018. *NBSL: A Supervised Classification Model of Pull Reques in GitHub.* Kansas City, IEEE International Conference on Communications (ICC).

Managing Large-Scale Heterogeneous Deployments for Cybersecurity

J.S. Hurley
ODNI, Bethesda, USA

john.s.hurley@odni.gov

Abstract: In cyber defense, we must contend with the massive amounts of data being generated in a variety of different formats and speeds. Unfortunately, traditional tools and methods are not meeting the requirements for scale and speed and rely too heavily on heuristics. Advancements in mobile technologies and the Internet of Things (IoTs) will continue to contribute to the additional growth in data volumes anticipated for the foreseeable future. As data continues to grow in complexity and scale, cyber professionals must rely upon models that are more elaborate and sophisticated to predict future behavior. More complex models can give additional inference capabilities; however, they are also difficult to scale and deploy in real-time environments. Managing large-scale, heterogeneous deployments for cybersecurity is challenging. Hardware capabilities and software tools both motivate and limit computational and inferential objectives. Hence, the interplay between data science (especially machine learning) and computation become more significant than ever to explore to gain more insight into heterogeneous deployments and how they can be more effectively managed. In this study, we identify ways in which data science tools and techniques can be used in improving the management of large-scale heterogeneous deployments for cybersecurity.

Key Words: Heterogeneous Deployments, Data Science, Heuristics, Inference, Cybersecurity.

1. Introduction

A large-scale set of heterogeneous devices reflects a fundamentally complex system with unique requirements and challenges driven by diverse devices with different operating systems and/or protocols. These deployments are complex, not only because of the need to integrate different languages, platforms, users, and technologies, but to do so seamlessly, with limited or no disruption of services. Other key requirements include the need to have continuous, resilient operations and high availability. Deployments must be flexible and dynamic enough to accommodate applications that can run in any environment. The ability to deploy at the edge, on-premises, or in the cloud requires diverse operation and security requirements that can complicate deployment. The advancements of technology, especially in the areas of mobile technologies and the Internet of Things (IoTs), demands that environments be readily scalable to accommodate future data growth. Similarly, network operators should anticipate that new systems and platforms will expand system diversity as emerging technologies and novel devices are incorporated into already complex systems.

It is critical to understand the layout of how devices and nodes are physically or logically connected and to understand how data will move through the network environment and where particular chokepoints or potential points of obstruction could exist within a network. An understanding of the network topology is a primary means of establishing effective network management and monitoring. No single network topology approach is perfect or inherently better than any other.

From a security and operations standpoint, it is important to understand and prioritize applications and the data associated with them. A consistent problem that has plagued organizations is the inability or unwillingness to prioritize network infrastructure and data management, including even keeping appropriate records of the location of critical data. Such behavior has unnecessarily complicated the ability of organizations to assign their limited resources to the most mission critical information assets which should have the most stringent security and operations requirements. Consistently, organizations treat all applications with the same value leading to unnecessary waste of funds, especially in terms of storage and security requirements. All applications and data should not be treated the same, and the most mission critical applications and data should receive the most attention.

Scalability and security are dominant issues in heterogeneous deployment. Network virtualization can play an important role in both, as strategies are increasingly being employed to improve heterogeneous deployment methods. Virtualization provides many network benefits, especially by allowing multiple isolated virtual networks to share the same underlying physical infrastructure. Virtualization strategies that virtualize computing, processing, and networking enable virtual networks to be added and scaled and allow networks to be spun up more quickly in response to shifting business requirements. Faster service delivery, improved control,

and enhanced operational efficiency are all byproducts of the flexibility that network virtualization allows (English, 2022). However, a downside of the virtualization is the network sprawl created when network administrators overindulge in the creation of virtual networks. In these cases, excessive resource consumption and network complexity can result. In addition, as enterprises migrate increasingly to virtual networks, the impact of new architectures on resource consumption, resilience and security must be given consideration.

In this study, data science tools and technologies will be used to address some of the challenges associated with managing heterogeneous deployments. The study will begin with a discussion of the different nodes, devices, and links and the importance of knowing how they are connected physically and logically. Next, the focus is on the issues of application and data prioritization and their importance in terms of where the limited resources available to an organization should be allocated, especially for security purposes. Next, the emphasis is specifically on the challenges associated with each of the computing environments. In the following section, considerations will be given to different data science tools and technologies and how they may alleviate some of the problems associated with the unique computing environments. Specifically, the focus will be on identifying and prioritizing the appropriate applications and data, then setting up a hierarchy of mission critical applications and associated data, wherein the data associated with the level of priority is placed in the appropriate storage locations. Next, the computing environments (cloud, edge, and on-premises) that are best suited for the selected applications will be coordinated. Lastly, overall recommendations that may mitigate some of the current challenges for deployment will be provided.

2. Network Topology

Network configuration or topology is key to determining network performance and provides the confidence to securely operate within the network. Information Technology (IT) and security professionals must have a keen understanding of how the different nodes, devices and links on networks are logically related and physically arranged to optimize network performance while mitigating network attacks. There are several options available to professionals with the usual constraints, i.e., size, scale, budget, mission, and goals of the organization. As expected, there are advantages and disadvantages to each option, but certain arrangements can provide a greater degree of security and connectivity. In physical network topology, there is greater emphasis on the physical connections and interconnections between the nodes and the network. The concept of logical network technology is a bit more abstract, referring to how the network is arranged as well as how data moves through the network. Both the physical and logical approaches to network technology play important roles in application and network performance.

3. Application/Data Prioritization

Organizations consistently engage in practices that adversely impact data, system and network performance and security. There is a general lack of knowledge of the level of priority of the applications and data for which an organization is responsible. In other words, applications, and data, for the most part are lumped into the same barrel with little to no distinction between those that are and are not mission critical. In addition, there is a general unawareness of what data the organization possesses; where the data is located; and how the data should be grouped or categorized. As a result, organizations tend to treat all data or at least most of the data with the same level of priority—an unnecessary use and cost of data storage and security resources. Not all data is created equally. It is more reasonable and practical for an organization to decide on the data that is sensitive or mission critical. Next, less critical data should also be assigned at the appropriate levels. All of this begins with first identifying and cataloguing the data to determine the types of data that the organization possesses. Finally, it is necessary to determine which individuals, groups, departments, and partners need certain data, and for what length of time.

4. Cloud, Edge, On-Premises Environments

We are familiar with on-premises and cloud computer environments. On-premises simply refers to an environment in which resources are deployed in-house and within an enterprise's local IT infrastructure. The services are generally privately owned and controlled locally. In a cloud environment, the resources are generally hosted on a vendor's server and not physically located in a local facility. In an edge computing environment, applications are brought closer to the data sources, which has numerous advantages, including improved response times and better bandwidth availability. A closer examination of the three different environments is made with distinct advantages and disadvantages of each being provided below.

An On-premises environment consists of IT infrastructure software and hardware applications that are hosted on-site. In spite of the local management and access to resources, on-premises environments face several challenges, including peak-capacity planning, continuous upgrades, reliability and runtime, high-availability, and disaster recovery. A major concern for an on-premises computer-aided design (CAD) infrastructure is reliability. Depending on the tracking methods, tracking, and calculating 100% uptime can be imprecise *(Varde, et al., 2021)*. On-Premises computing is about addressing computing requirements on the site of the organization. The biggest advantage likely for on-premises computing is security and data protection, since the data is stored locally on site, so there is full control over the data and the security. Locally stored data can still be accessed on the premises even when external network or Internet connections are interrupted. Upfront costs can be very high given that the functionality of the programs and back up of the data must be ensured. Typically, hardware requirements must be met by on-site devices because the software runs locally which can contribute to the costs of maintaining the right hardware. Scalability of on-premises environments remains a challenge from a cost and infrastructure reframing standpoint *(Kemper, 2021)*.

In cloud computing, the focus is on the on-demand access to services and resources via the Internet that are usually hosted at a remote data center that is managed by a cloud service provider (Srivastava & Khan, 2023). *For cloud computing, there are usually three distinct categories defined including: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Some of the advantages of cloud computing are Reliability, Strategic Edge, Security (depending on who you ask. Security has been improved, but there is still a high reluctance depending upon the cloud environment chosen. Private and hybrid cloud environments due to compliance requirements for certain organizations and applications are the only options available* (Duggal, 2022). A core component of the cloud-native computing environment, microservices, is being widely embraced as a very new and popular approach for developing and deploying cloud applications that require higher levels of agility, scalability, and reliability. A microservice-based cloud application architecture advocates decomposition of monolithic application components into "microservices" – independent software components. Unfortunately, this also creates complex runtime performance monitoring and management challenges because the independent microservices can be developed, deployed, and updated independently of each other (Noor, et al., 2019). The strength of the microservice model (agility, independence, diversity) also presents cybersecurity and monitoring challenges.

Edge devices are associated with hardware that controls data flow at the boundary between two networks. These devices can function in a variety of different ways including filtering, monitoring, processing, routing, storing, translating, and transmitting data between networks. In edge computing, processes are decentralized and occur in a more logical geographic location. The IoTs and cloud computing have raised the profile of edge devices, echoing the need for more advanced services, computing power and intelligence at the edge of the network. Unfortunately, Edge devices can increase the risk of cybersecurity threats to an organization's network. As might be expected, the deployment of hundreds of edge computing devices creates hundreds of potential entry points for Distribute Denial of Service (DDoS) attacks and other security breaches. The level of vulnerabilities is especially concerning from a security perspective because many of the endpoints are "smart," i.e., features are built-in for Internet connectivity. In addition, IT no longer has full visibility or centralized control. CISOs and CTOs are then presented with the significant security challenge of protecting the data that moves through or resides in edge devices, as the attack surface expands *(Ray, 2022)*. The capacity of edge devices is much more modest than that of a mini-cloud, and as a result requires new algorithms, methods, and policies to set the proper virtualization strategies for edge devices. Edge computing enables data to be analyzed locally, i.e., closer to where the data resides, in or near real time with minimum latency. Some advantages of edge computing are high speed, reduced latency, better reliability for faster content delivery and data processing. Edge computing offers a far less expense for versatility and scalability. Some of the disadvantages of edge computing are more storage capacity required, security challenges due to high amount of data. Edge computing also requires advanced infrastructure *(Arora, 2022)*.

5. Results and Discussions

In this study, we seek to address a number of the critical issues that plague organizations in terms of their applications and data in heterogeneous deployments. It is concerning how few organizations can lay (digital) hands on or even describe the location of their critical data when needed. Equally disturbing is the unawareness of how the data is being accessed and with whom it is shared. Most organizations still have

trouble addressing these issues and as such are significantly challenged by new global compliance requirements that mandate the understanding of the location of sensitive data as well as the level of security and protection of that data. The fact that data growth is expected to continue in the foreseeable future suggests the problem will only get worse as time goes on. With growing requirements and limited resources, we suggest a nuanced approach where all data is not treated the same. Given the limited resources within our organizations, it is critical that we establish hierarchies from the highest to the least critical data. It is evident that data that is critical to the Mission of the organization should receive the highest priorities. Hence, within this study, the attempt is made to understand where applications and data can be located so that data can be accessed more quickly. In this study, there are ten applications identified from E1-E10, with each having an arbitrary focus area that ultimately decides their position due to how critical the application and data are to the Mission of the organization, see Table 1. The requirements for the applications were generated randomly. The data for each of the applications is stored in a specific storage location, similarly, identified as storage data location (SDLn) from S1-S10. The applications are later ranked in terms of how critical they are to the Mission of the organization from High to Low, see Table 2.

Table 1: Applications, Focus Areas, Mission Critical Status, Storage Data Locatio

AppName	Focus Area	MCS	SDLn
E1	FA1	Medium	S1
E2	FA2	Low	S2
E3	FA3	High	S3
E4	FA4	Low	S4
E5	FA5	Medium	S5
E6	FA6	Medium	S6
E7	FA7	High	S7
E8	FA8	High	S8
E9	FA9	Low	S9
E10	FA10	Medium	S10

Table 2. Applications, Focus Areas, Mission Critical Status, Storage Data Location—Prioritized by MCS

AppName	Focus Area	MCS	SDLn
E3	FA3	High	S3
E7	FA7	High	S7
E8	FA8	High	S8
E1	FA1	Medium	S1
E5	FA5	Medium	S5
E6	FA6	Medium	S6
E10	FA10	Medium	S10
E2	FA2	Low	S2
E4	FA4	Low	S4
E9	FA9	Low	S9

In Tables 3. And 4., the emphasis is placed on criteria for the three different computing environments of note, including the cloud, edge, and on-premises environment. In Table 4., the environment criteria in Table 3. is converted from categorical to quantitative representations to employ in computing program.

Table 3. Computing Environments Criteria.

Cloud	Edge	On-Premises
Little to No Upfront Costs	Most Cost-Effective	Large Upfront Costs
Data Ownership Not Transparent	Better Data Sovereignty than Cloud	Complete Control of Data
High Latency	Lowest Latency and Congestion Solution	Low Latency
Some Cybersecurity Concerns	Potential Data Privacy and Cybersecurity Concerns	Stronger Control of Security
Lower Than On-Premises Real Time Performance	Highest Real Time Performance Advantage	High Real Time Performance Advantage
Highest Scalability	Higher Scalability than On-Premises	Least Flexibility
Highest Bandwidth Flexible	Higher Bandwidth Flexibility than On-Premises	Bandwidth Upgrade Capability Lowest
Lowest Control Host, Management, and Some Maintenance	Higher Host, Management, Maintenance Control than Cloud	Highest Level of Control
Reliable, but Internet Connection Critical	Higher Liability than Cloud	Highest Reliability
Database or Performance-Intensive Apps	Greater Preferred than Cloud	Most Preferred

Table 4. Computer Environments Criteria (Numerical Conversion)

Requirements	Cloud	Edge	On-Premises
Costs	2	3	1
DO	1	2	3
Lat	1	3	2
Cybersec	1	2	3
RTP	1	3	2
Scale	3	2	1
BW	3	2	1
HMM	1	2	3
Rel	1	2	3
Apps_PI	1	2	3

Legend: Costs = Costs; DO=Data Ownership; Lat=Latency; Cybersec=Cybersecurity; RTP=Real-Time Performance; Scale=Scalability; BW=Bandwidth; HMM=Host, Management, and Maintenance; Rel=Reliability; Apps_PI=Performance-intensive Apps. 1º lowest value; 2º medium value; 3º highest value.

In Table 5., a collective arrangement of the apps, criteria, and computing environments are listed. In Figures 1., 2., and 3. correlation and heatmap plots for the applications, criteria and computing environments are presented, respectively.

Table 5. Full Apps and Computer Environments

Apps	Costs	DO	Lat	Cybersec	RTP	Scal	BW	HMM	Rel	Apps_PI	Cloud	Edge	On_Premises
1	3	3	3	1	3	1	3	2	1	2	2	3	1
2	2	2	2	3	1	3	3	3	3	1	1	2	3
3	2	3	2	2	3	3	3	1	1	1	1	3	2
4	2	3	1	2	3	1	3	1	1	3	1	2	3
5	3	3	2	1	1	2	2	2	2	3	1	3	2
6	1	2	2	1	2	1	1	2	3	3	3	2	1
7	2	1	2	1	3	1	2	1	2	3	3	2	1
8	2	1	2	3	2	1	3	2	2	1	1	2	3
9	2	3	2	2	2	1	1	1	3	2	1	2	3
10	3	1	2	3	1	3	3	3	1	2	1	2	3

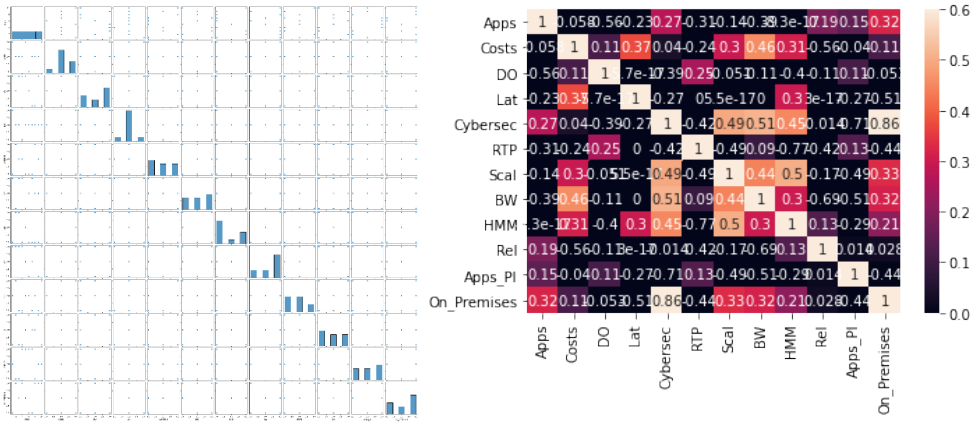


Figure 1: Correlation and Heatmap for on_Premises

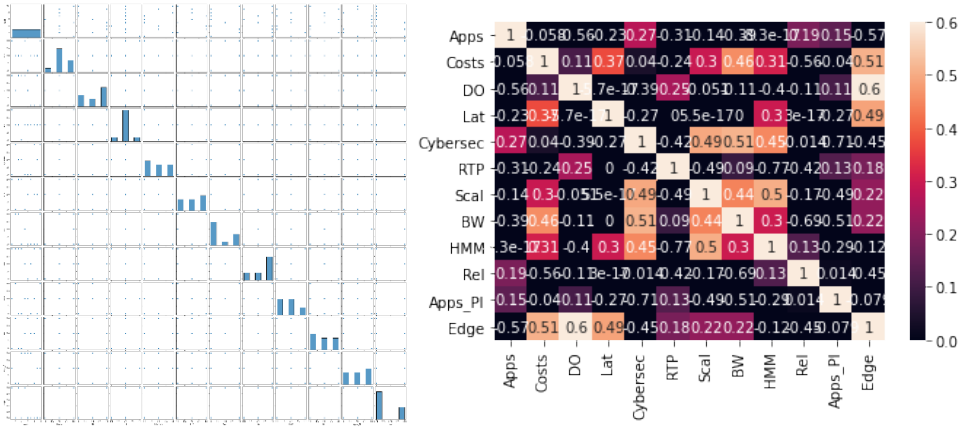


Figure 2: Correlation and Heatmap for Edge

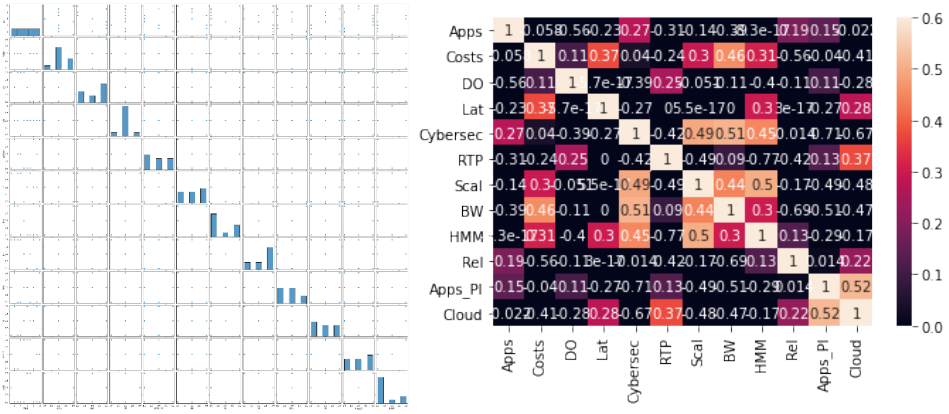


Figure 3: Correlation and Heatmap for Cloud

In this study, the Decision Tree algorithm was used to establish the relationships between the Applications (1-10), the criteria (see legend in Table 4.), and the three different computing environments, i.e., Cloud, Edge, and On_Premises. Training and testing data sets were split into a 75/25 ratio with the testing data sets being arbitrarily chosen, and in this case, they are Apps 3, 9, and 5. However, noticed that there was not much difference when considered at 60/40 training to test split. The goal was to predict the appropriate computing environment given the criteria established for the applications. The numbers for the criteria were generated randomly to recognize the disparity and differences in real world computing environments for different organizations depending on their size, resources, requirements, and Mission. The heatmap and correlation plots were selected to focus on the correlation between the parameters of interest. For the focus on cybersecurity, it is interesting to note that cybersecurity for performance-intensive apps for edge and cloud computing were

relatively close in correlation with a value of ~ 0.710. For the on-premises security environment, the correlation to the performance-intensive (including databases) apps was higher, to be expected, than the others with a value of ~0.860. There was an interesting observation with all three computing environments, wherein the RTP and HMM pairing yielded a value of ~ -0.77. Further analysis is required.

6. Conclusion

In this study, the focus was on addressing some of the major challenges associated with deploying applications within heterogeneous environments, including cloud, edge, and on-premises. A way to formally prioritize data and applications in terms of Mission importance was established. The efforts emphasized the relationship between Mission Critical applications, relevant data and storage locations, and the identification of the "right" compute resource for which an application is best suited. The applications, focus areas, and storage locations are all defined anonymously in this work and left to the determination of the organization. This work sought to prioritize data from top to bottom at different levels, define the location of the data in terms of importance—a continuing problem within many organizations that still do not know where a lot of their mission critical data is located. Knowledge of the most mission critical data to the least critical data allows organizations to more intelligently define and implement a budget for resources such as storage, as well as gain access to critical data when needed.

The correlation analysis was selected for this study to emphasize the similarity of variables of interest (especially as each variable relates to cybersecurity) when examining the criteria for defining the compute resource of choice. The correlations (positive and negative) are essential in reflecting how variables are trending in the same or opposite directions, respectively. The correlations were especially helpful in classifying applications according to their relevance to Mission. It is important to note that there will surely be different levels of the applications and data within the classes. For example, applications E3, E7, and E8 are each classified as high, but surely there is a hierarchy to the levels that can still easily be resolved through an additional application of the process until each application is addressed according to its Mission critical status from top to bottom. Future work will focus on defining a pipeline that allows the automation of the entire process, as well as, examining other machine learning algorithms, including an ensemble method to see how results compare between different algorithms.

References

Arora, S., 2022. *Edge Computing vs Cloud Computing: Key Differences to Know.* [Online]
Available at: https://www.simplilearn.com/edge-computing-vs-cloud-computing-article
[Accessed 19 01 2023].

Duggal, N., 2022. *Advantages and Disadvantages of Cloud Computing.* [Online]
Available at: https://www.simplilearn.com/advantages-and-disadvantages-of-cloud-computing-article
[Accessed 19 01 2023].

English, J., 2022. *What is Network Virtualization? Everything You Need to Know.* [Online]
Available at: https://www.techtarget.com/searchnetworking/What-is-network-virtualization-Everything-you-need-to-know#:~:text=External%20virtualization%20uses%20switches%2C%20adapters,without%20using%20an%20external%20network.
[Accessed 19 01 2023].

Kemper, F., 2021. *On-Premises vs Cloud: Advantages and Disasdvantages.* [Online]
Available at: https://www.empowersuite.com/en/blog/on-premise-vs-cloud
[Accessed 19 01 2023].

Noor, A. et al., 2019. *A Framework for Monitoring Microservice-Oriented Cloud Applications in Heterogeneous Virtualization Environments.* Milan, IEEE.

Ray, M., 2022. *Edge Computing: The Advantages and Disadvantages.* [Online]
Available at: https://innovatorscentral.ca/edge-computing-advantages-and-disadvantages/
[Accessed 12 01 2023].

Srivastava, P. & Khan, R., 2023. A Review Paper on Cloud Computing. *International Journals of Advanced Research in Computer Science and Software Engineering,* 8(6), pp. 17-20.

Varde, A., Bhonge, N., Duffield, M. & Morris, S. M., 2021. *Best Practices for Developing ClioSoft SOS on AWS,* Seattle: Amazon .

Digital Geopolitics: A Review of the Current State

Gazmend Huskaj
Geneva Centre for Security Policy, Geneva, Switzerland

g.huskaj@gcsp.ch

Abstract: The purpose of this research product is to present the current state of digital geopolitics. Digital Geopolitics is attracting much attention. It features in national digital strategies (for those countries that have those), and there is some research on the topic. However, until now, no systematic and up-to-date review of the scientific literature on digital geopolitics exists. This article reviews the scientific literature using the computational literature review method. 124 articles were identified in a scientific database. After removing articles without author and abstract, 120 articles remained to read, cluster and present in this research product. The findings present that research output increases from 2015 and onwards, 53 topics are covered in the data set, and top cited articles and top publication venues are presented. The answer to the research question is that based on the results and the manual clustering of topics, it is indicative that the Technology, Informational, Geography are Security areas have a high focus, with less focus on, for example, political and health areas.

Keywords: current; state; digital; geopolitics; review

1. Introduction - Why do we have to talk about Digital Geopolitics?

This article reviews the question what the current state of digital geopolitics is. Why is there a need to discuss Digital Geopolitics? There is no technology that has transformed societies in the ways that information- and telecommunications technology (ITC) has. The increased pace of digitalisation has opened opportunities to conduct e-commerce, e-governance, streamline processes, but also transformed intelligence, surveillance, reconnaissance, and command and control of joint military operations. The increased interconnectedness has also shortened the gap: threat actors can conduct successful attacks from their own territory, posing a threat to political stability, military secrets, and economic and social well-being. Cases such as the ransomware attacks on the private actors Colonial pipeline and Kaseya, and cyber espionage attacks on English-speaking organisations, dubbed APT1 (Hyvärinen, 2017), highlight how attacks like these can impact political, economic, social, and international stability.

Geopolitics is the "analysis of geographic influences on power relationships in international relations" (Deudney, 2013). The Swedish political scientist, Rudolf Kjellén, coined the term by the end of the 19th century or the beginning of the 20th century (Deudney, 2013). Kjellén argued that geopolitics is about "the problems and conditions within a state that arise from its geographic features" (Albert, 2022). Discussions on the political effects of geography on state security is nothing new. They range from Aristotle (384-322 BC), to Mackinder (1861-1947) (Deudney, 2013). Examples include climate, topography, and by the end of the 19th century, and beginning of the 20th century, the impact of new technologies on world politics. First, it was argued that control of the seas was important due to trade and transportation, that is, sea power. Then, when new technological advances led to railroads, new arguments were made that trade and transportation was now faster than sea routes, why land power was more important than sea power. With the discovery of the aeroplane, trade, transportation, and ways of waging war were different. Now, some geopoliticians and generals, like Giulio Douhet, argued that air superiority and air power theory was more advantageous than sea and land (Deudney, 2013). In summary, geopoliticians tried to understand how new technologies and related capabilities, during their interaction "with the largest-scale geographic features of the Earth would shape the character, number, and location of viable security units in the emerging global international system" (Deudney, 2013).

Digital, an adjective, is from the mid 15th century, "pertaining to numbers below ten" (Online Etymology Dictionary [OED], (n.d.)), from Latin digitus "finger or toe", where the "numerical sense is because numerals under 10 were counted on fingers" (OED, n.d.). During the development of the computer, digital could mean "using numerical digits" (from 1938); computers running on data in the form of digits (from 1945), and "in reference to recording or broadcasting, from 1960" (OED, n.d.). Digital, thus, is about using computers to compute data in the form of digits (often ones and zeroes). Today, in addition to digital, digitisation, digitalisation and digital transformation exist. According to Bloomberg (2018), digitisation "refers to taking analog information and encoding it into zeroes and ones so that computers can store, process, and transmit such information" (Bloomberg, 2018). Digitalisation however is not equally straightforward because many different definitions exist. One is about "the way in which many domains of social life are restructured around digital communication and media infrastructures" (Brennen & Kreiss, as cited by Bloomberg, 2018). Another states that digitalisation "is the use of digital technologies to change a business model and provide new revenue and value-producing

opportunities" (Gartner, as cited by Bloomberg, 2018). A third states that "digitalisation is the process of employing digital technologies and information to transform business operations" (Gartner, as cited by Muro et al. (2017), as cited by Bloomberg, 2018). In summary, digitalisation is the use of information systems (computers) and information- and telecommunication systems (ICT) to not only restructure societies, but also to restructure business operations. Thus, digital transformation is about the transformation of key business operations that affect products, processes, organisational structures, and management concepts, with the purpose of exploiting the benefits of digital technologies (Matt, Hess & Benlian, 2015).

In this research product, Digital Geopolitics is defined as the analysis of information systems (computers that create, process, store, retrieve and disseminate information, [Huskaj, 2019]) interconnected in networks of networks to build a large, global network (Cyberspace [Huskaj, 2019]), that together with the geographic features of Earth and space impact world politics, the military, economic, social, informational and infrastructure (PMESII) systems and subsystems. This is also why it is important to discuss digital geopolitics: cyberspace is ubiquitous and affects all domains of human life. Dunnett et al. (2017) provide discussions on the geographies of outer space.

2. Methods and Materials

Systematic literature reviews fulfil the criteria of "objectivity" and "repeatability". The computational literature review (CLR) method uses topic modelling based on the latent Dirichlet allocation algorithm (Blei et al., 2003). Additional algorithms were used in this research product, these are presented further down.

The source for collecting research products on digital geopolitics was Scopus. The Scopus® database is, 'the largest abstract and citation database of peer-reviewed literature' (Elsevier 2018). Scopus is a database that contains curated research information in a specific format. This makes it ideal to download and process the information computationally. On June 17, 2022, the keywords digital geopolitics were used in article title, abstract and keywords. The database presented 124 document results, from 1999 to 2022, published in 104 publication venues. The documents were cleaned to remove documents without author, abstract, or doubles, leaving 120 articles for review.

A first review of the 120 articles using the CLR-method revealed few insights and the decision to make a granular review of a subset of the dataset between 2015 to 2022 was taken. The 99 documents were cleaned by removing articles without author(s) and abstract, leaving 96 articles for review. The dataset presents an increase of research products per year, as Figure 1 and Figure 2 presents.

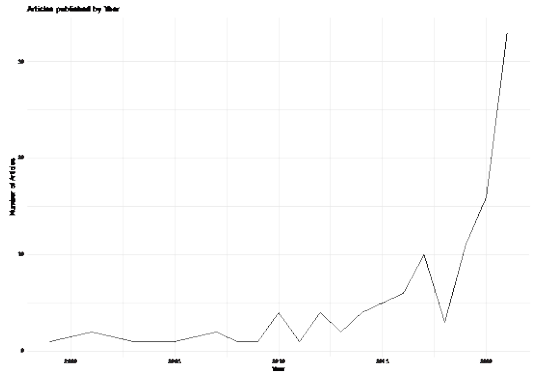


Figure 1: Research products published per year from 1999 to 2021.

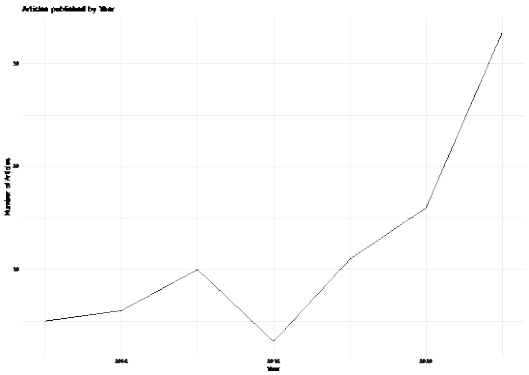


Figure 2: Research products published per year from 2015 to 2021.

Two additional algorithms were used in this research product. The ldatuning-algorithm (Nikita, 2016), and a modified algorithm inspired by the correlated topic model-algorithm (Blei & Lafferty, 2007) used by Roberts, Stewart, Tingley, and Airoldi (2013). The correlated topic model (stm) also includes information about trends over time, which makes it possible to discern how the interest in the different identified topics change over time. The keen reader will most likely ask the question how do you know how many topics reside in a dataset? Previously, researchers had to pre-set a value for the number of topics in a dataset and iterate the process until the results provided a set of topics which seemed reasonable. The topic-values were previously set to 30, 60, 90, or 120. However, today, the ldatuning-algorithm supports the research analyst in the relevant topics in a

dataset, also known as the k-value. Figure 3 presents the relevant topics, the k-value, for the dataset on digital geopolitics from 2015 to 2022.

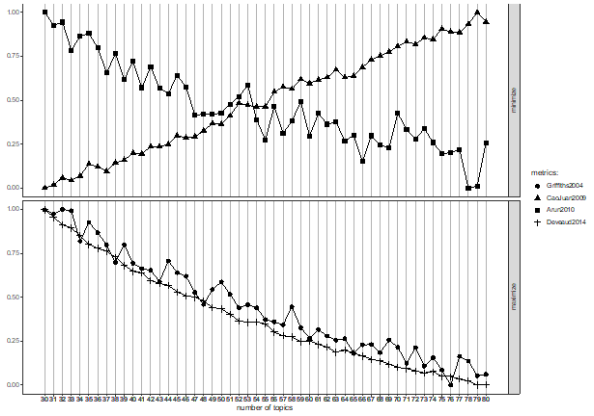


Figure 3: The ldatuning-algorithm to identify the number of topics in the dataset. The x-axis depicts no. of topics, the y-axis depicts algorithm-metrics.

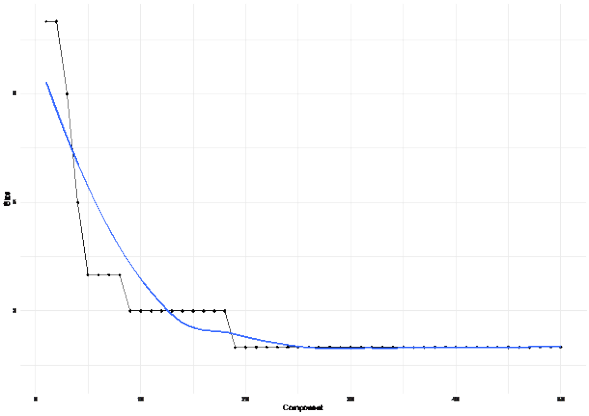


Figure 4: Presenting the 50 first co-author networks, where the top three consist of 11 co-authors, 11 co-authors, and 9 co-authors.

The results of the ldatuning-algorithm present that 53 topics exist in the dataset. Both algorithms, the LDA and the CTM, may be used to present each topic in word clouds. Word clouds is a method to visually present text. They present differences, similarities, and which words are most likely in a topic. The colours show different categories, and the size of the font presents the likelihood of which words will occur. An example of a word cloud is presented in Figure 4. It is up to the research analyst to infer the word clouds and provide a relevant label for the topic they discuss.

Finally, each topic consists of a set of research documents. These documents are, through the application of the LDA or CTM, statistically clustered into topics. Therefore, each topic, labelled by the research analyst, can be further analysed to present which research documents exist within that topic. Table 3 presents the insights provided by the 53 topics.



Figure 5: Topic 6 discusses geopolitical risk.

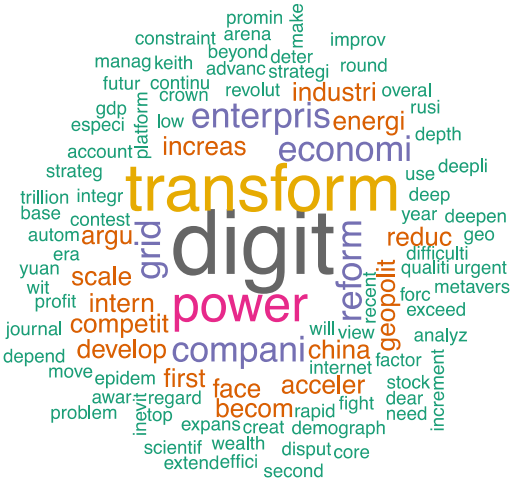


Figure 6: Topic 10 (the right) discusses digital transformation.

Finally, to cover some additional areas, the following five questions were used:

How does Digital Geopolitics relate to PMESII?; What factors can impact the geopolitical order and balance of power?; How is digital geopolitics important for the future of the Internet?

3. Results

The topic modelling presents 53 topics in the dataset. The topics are evaluated by the word clouds. The word clouds, topics, and papers, are presented in the supplementary material. Table 1 presents the most cited articles, while Table 2 presents the publication venues. The co-author of the authors is presented in Figure 5. Finally, as mentioned above, Table 3 presents the qualitative analysis of the 53 topics, manually inferred and labelled by the research analyst. Figures 6 to 8 present trends over time.

Table 1: Top 5 articles ranked by citation count.

No.	Authors	Title	Year	Source	Cites
1	Aouragh M., Chakravartty P.	Infrastructures of empire: towards a critical geopolitics of media and information studies	2016	Media, Culture and Society	44
2	Yang S., Yu X., Ding J., Zhang F., Wang F., Ma Y.	A review of water issues research in Central Asia	2017	Dili Xuebao/Acta Geographica Sinica	23
3	Peckham R., Sinha R.	Satellites and the New War on Infection: Tracking Ebola in West Africa	2017	Geoforum	18
4	Lavin M., Yang L., Zhao J.J.	Boys' love, cosplay, and androgynous idols: Queer fan cultures in mainland China, Hong Kong, and Taiwan	2017	Boys' Love, Cosplay, and Androgynous Idols: Queer Fan Cultures in Mainland China, Hong Kong, and Taiwan	17
5	Ermoshina K., Musiani F.	Migrating servers, elusive users: Reconfigurations of the Russian internet in the post-snowden era	2017	Media and Communication	15

Table 2: Top 5 publication venues ranked by number of articles and citation count.

Rank	Venue	Articles	Rank	Venue	Cites
1	Geopolitics	4	1	Media, Culture and Society	51
2	Media, Culture and Society	2	2	Dili Xuebao/Acta Geographica Sinica	23
3	Geoforum	2	3	Geoforum	22
4	Internet Policy Review	2	4	Geopolitics	16
5	Environment and Planning D: Society and Space	2	5	Media and Communication	15

Table 3: The 53 topics identified by the algorithm.

Topic no.	Insights
1. Natural Gas	Discusses natural gas supplies in China and related consumption (1.1).
2. Digitalisation and the Internet	Discusses digital diplomacy (2.1), disinformation (2.2) and the impact of technology (2.3).
3. Internet Geopolitics	Impacts on Indian cyber security (3.1), digital challenges (3.2), Internet Diplomacy (3.3).
4. Insecurity	Insecurity and security in young people (4.2) and implications for trade (4.1).
5. Definitions and Communities	Definitions on cyberwar (5.2) and social media's impact on community development (5.1).
6. Geopolitical Risk	Geopolitics and technological risks (6.1, 6.2).
7. The three seas initiative	A model for regional cooperation in central Europe (7.1).
8. Different digital discourse and tactics	Discourse analysis on the Internet (8.1) and tactics to be elusive (8.2).
9. Games for geopolitics	Visual novels (9.1), strategy games for education (9.2), and games for geopolitics (9.3).
10. Digital transformation	Digital transformation to empower power grids (10.1), and how digital transformation is impacting on power in geopolitics (10.2).

Topic no.	Insights
11. Digital war	The U.S.-China tech war (11.3), technology and infrastructure (11.2), and cultural geopolitics (11.1).
12. Geopolitics of travel blogging	Geopolitics and travel blogging (12.1).
13. Game layers	Game images and tech obsolescence (13.1).
14. Digital economies	Digital capacity of Russian tourist territories (14.1), and China in international digital economy governance (15.1).
15. Transition to the Cybersphere	China's transition to the cybersphere and its temporal-spatial governance (15.1).
16. Geographical places	Mediated geographies (16.1) and exploring the semantic and geographic space of Catalonia (16.2).
17. The arctic	The arctic and future opportunities in logistics (17.1).
18. Undersea Cables	The security politics of undersea cables (18.1).
19. European Energy Imports	European Energy import demands (19.1).
20. Geopolitics and digital power competition	The geopolitics of social media platforms (e.g. TikTok) (20.1), the impact of geotechnics (automation, machines) on geopolitics (20.2), Geopolitics, jurisdiction and surveillance (20.3).
21. Chinese interests	Tech-geopolitics (US-China trade war) (21.1), and Huawei's expansion (21.2).
22. Digital mapping	Using digital maps in disputes (22.1), mapping North Korea (22.2), and the Geographies of digital culture (22.3).
23. Cybersecurity	Cybersecurity development in the context of digital transformation (23.1), and the EU's response to cyber threats (23.2).
24. Digital geopolitics	Digital geopolitics of grime (24.1), and geopolitics and digital representations of space (24.2).
25. Data colonialism	Decolonising data colonialism (25.1).
26. Employment and platforms	Platforms of work, labour and employment (26.1), and digital relationships (26.2).
27. Education	A "Covid Collective" - the Philosophy of education (27.1).
28. Digital performance	Online performance (28.1) and the digital humor of the baltic Russian-speaking social media users (28.2).
29. States and algorithms	Cyberspace is ungoverned (29.1) and proactive state geographies (29.2).
30. Maps and claims	The role of maps in territorial disputes (30.1).
31. Data infrastructure	The Schengen Information System (31.1).
32. Fan cultures	Queer fan cultures in China, Hong Kong and Taiwan (32.1), and dismantling authoritarianism in Asia through digital means (32.2).
33. Foreign interference	Attacking democracy through digital means (33.1).
34. Unidentified	IP, piracy and counterfeiting (34.1).
35. Freedom of Expression	Freedom of expression in digital transition (35.1).
36. Education and humanoid robots	Humanoids in digital health education (36.1).
37. Connectivity and economic growth	Alternatives to connectivity (37.1), economic growth, and war on terror (37.1).
38. Dissent	Conflict and visual politics (38.1), and dissent in the digital age (38.2).
39. Humans, States and Satellites	Human rights through visual culture (39.1), state power and technology in India (39.2), satellite-based activism (39.3) and satellites to track Ebola (39.4).
40. Soft Power	Soft power and diplomacy in the pacific (40.1).
41. Vaccination Themes	The Sputnik vaccine (41.1) and digital vaccine certification (41.2).
42. Platforms	The globalisation of TikTok (42.1) and countering platform finance (42.2).
43. Processes	Commodification processes in relation to knowledge and publishing in journals (43.1) and triunity in the process of formation (43.2).

Topic no.	Insights
44. Migrant caravan	The digital life of #migrantcaravan and Twitter as a spatial technology (44.1), and aspects within the gulag architectonic (44.2).
45. Big data	The role of big geospatial data in the Covid-19-pandemic (45.1), big data and ethical challenges (45.2), and a cross-disciplinary exploration on AI-literacy (45.3).
46. Global internet	Traffic analysis in Cyberspace (46.1) and Chinese thoughts on cyber sovereignty (46.2).
47. Water in Central Asia	Water issues in Central Asia (47.1).
48. Digital development	The three seas initiative, geopolitics, economics and infrastructure (48.1, 48.2), and Russia's geodigital interests (48.3).
49. Business relationships	Australian-European economic and geopolitical relationships (49.1).
50. Copper logistics	Logistics along the line of copper (50.1).
51. Global digital currency	Digital currency (stable coin) as a reserve asset (51.1) and digital communication disrupting hegemonic power (51.2).
52. Digital storytelling	Digital storytelling (52.1, 52.3), and the geopolitics of Digital Space (52.2).
53. Digital platforms	China's stance against tech monopolies (53.1).

Manual clustering the topics reveals the following areas as depicted in Table 4.

Table 4: Manual clustering of topics into the different areas. NOTE: each number represents the topic number as depicted in Table 3.

Domain	Sub-topic	Domain (cont.)	Sub-topic (cont.)
Political	40	Energy	1, 19
Military	5, 11	Technology	2, 3, 10, 13, 15, 20, 24, 25, 28, 29, 31, 39, 42, 48, 51, 53
Economic	14, 37, 49	Education	37, 36
Social	9, 26, 32, 44	Health	41
Informational	8, 12, 22, 35, 38, 43, 45, 52	Security	4, 6, 23, 33, 47
Infrastructure	7, 46, 50	Unidentified	34
Geography	16, 17, 18, 21, 30		

In the introduction, the case for why digital geopolitics is important to discuss was made: cyberspace is ubiquitous, affects all domains of human life, and supporting PMESII-systems. This section will present how digital geopolitics relates to PMESII.

3.1 How does digital geopolitics relate to PMESII?

The ransomware attack, by the Russian-based DarkSide, on Colonial pipeline, the largest fuel pipeline in the U.S., had a great impact on the daily lives of Americans (Turton & Mehrotra, 2021). First, because the attackers encrypted the information systems of Colonial, the company was unable to operate the pipeline. Colonial decided to turn off the pipeline, the first time since it was turned on in 1964 (Turton & Mehrotra, 2021). The cyber security firm Mandiant was called in to investigate the attack (Kerner, 2022). While the pipeline was turned off, gas stations were unable to receive the standard amount of gas, which led to some gas stations running out. There were also reports that this could hit the aviation industry. The impact on the ground was long queues of Americans to buy gas, where the price of gas rose to $3.05 per gallon (~ 4.55 litres) (Dean, 2021). The U.S. government deemed the ransomware attack a national security threat and the President declared a state of emergency (Kerner, 2022). Management paid the $4.4 million ransom (Turton & Mehrotra, 2021). The FBI was able to follow the trace of bitcoin and recovered around $2.4 million after the Department of Justice "got a court order to seize the bitcoin" (Kerner, 2022). On June 16, at the Biden-Putin summit in Geneva, Switzerland, President Biden informs President Putin that some critical infrastructures should be "off-limits" (Soldatkin & Pamuk, 2021). Later, the partner of DarkSide, REvil, was targeted by the FBI, US Cyber Command (CYBERCOM), the Secret Service, and like-minded countries (Kellerman, as cited by Menn & Bing, 2021).

Analysing the case results in the following stakeholders involved: DarkSide and the target, Colonial Pipeline. Those impacted were suppliers of fuel and the citizen. The US President declared a state of emergency, the DoJ got a court order, the FBI recovered some of the ransom, and President Biden and President Putin met in Geneva.

Later, the FBI, US CYBERCOM, the Secret Service and like-minded countries were involved, dismantling REvil's infrastructure. From a PMESII-systems-perspective, the ransomware attack impacted world politics (President Biden and President Putin met in Geneva), CYBERCOM (military) conducted offensive cyberspace operations on REvil's infrastructure, the gas prices (economy) went up, long queues to gas stations developed (social), Colonial pipeline's information systems (information) were taken ransom, Colonial pipeline is critical infrastructure (infrastructure), finally, the attack was done by a Russian-based threat actor (geography). The remaining two cases, Kaseya and APT1, are left to the reader to analyse according to the case above.

3.2 So, what factors can impact the geopolitical order and balance of power?

As noted above, in Table 4, any impact on geographical, space, and PMESII-systems and subsystems can affect the geopolitical order and balance of power. One area that is overly presented is Technology. Advances in technology have, as mentioned in the introduction, shaped human life, societies and even the geopolitical order and balance of power. Topic 11 for example, discusses the U.S.-China Tech War, and Topic 21 discusses tech-geopolitics and Huawei's expansion. The U.S.-China tech war is a complicated topic. However, the Chinese companies Huawei and ZTE provide numerous technologies, where 5G telecommunications being one of them. 5G telecommunications equipment is needed to build the backbone of 5G networks. However, the U.S. intelligence community, and their counterparts in the U.K., Australia, Canada, and Sweden, to name a few, have voiced Chinese e-espionage-concerns of implementing 5G telecommunication equipment from Huawei or ZTE in the backbone infrastructure for 5G networks. The Chinese Communist Party (CCP) have accused the U.S. of protectionism. These back-and-forth accusations have led to what journalists call a "trade war".

Topic 2, 33, and 35, cover informational topics such as disinformation, foreign interference, and freedom of expression. Disinformation is extensively used by Chinese- and Russian-based threat actors, but also criminal entities. This has been especially evident during the Covid-19-pandemic, where Chinese-based threat actors have conducted cyber espionage on labs developing vaccines (BBC, 2020); Russian-based threat actors conducted cyber espionage on labs developing Covid-19-vaccines (NCSC, 2020). In addition, Covid-19-related disinformation campaigns from China, Iran, Russia and Tunisia providing false information have been identified and taken down by Google (Cimpanu, 2020). The combination of cyber espionage on labs developing Covid-19-vaccines and disinformation campaigns providing false Covid-19-information, lead to the Chinese and Russians accelerating their own Covid-19-vaccines at a cheaper cost (they're stealing the knowledge), at the same time, through disinformation campaigns erode trust between populations and their own governments, leading to phenomena as "anti-vaxxers".

In summary, as noted from the results presented in Table 3 (insights), and real-world cases, it is evident that factors like geography, space, and PMESII-systems, can impact the geopolitical order and balance of power.

3.3 Now, how is digital geopolitics important for the future of the Internet?

The analysis of internetworked information systems together with the geographic features of Earth and space are all important to monitor with the purpose of identifying threats, vulnerabilities, and related risks to the same. The purpose is to ensure that information that is created, processed, stored, retrieved and disseminated by information systems and humans, remains confidential, has its integrity untouched, and is available to those that have the right to access that information. It is evident that as technology has had huge positive transformational impacts on societies, the same can be misused by threat actors, ranging from states, state-sponsored threat actors, to organised crime groups and lone-wolf actors. Thus, as 5G-technologies are implemented, 6G-technology-seminars are held, to ensure that societies still thrive from these technologies, it is important that current and future leaders, professionals and technologists, work to ensure a safe, secure, and trustworthy Internet.

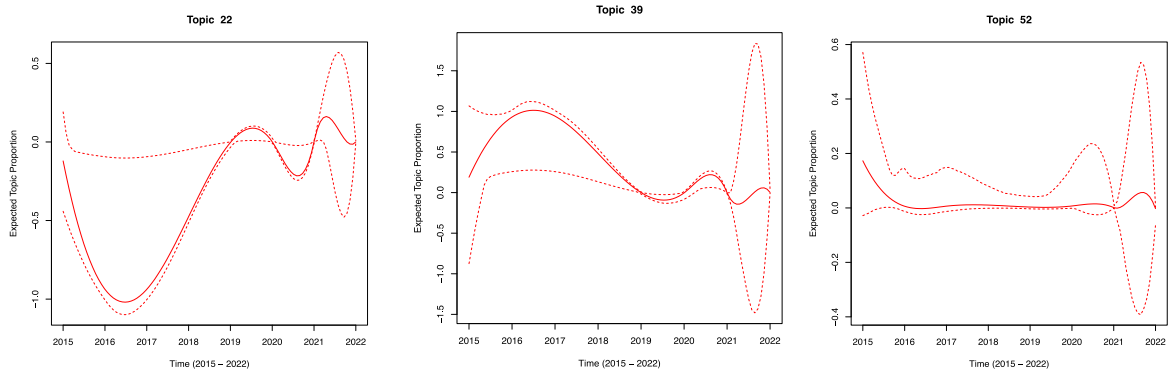


Figure 7: Trends on three topics from 2015 to 2022. Topic 22 – How digital maps are increasingly used in territorial disputes; Topic 39 – A declining trend about human rights through visual culture, state power and technology in India, and satellite-based activism; Topic 52 – How digital storytelling has shifted during the years.

4. Discussion

The answer to the question *what the current state of digital geopolitics is,* is that based on the results and the manual clustering of topics, it is indicative that the Technology, Informational, Geography are Security areas have a high focus, with less focus on, for example, political and health areas. Table 5 presents the biggest area-cluster in a descending order.

Table 5: The biggest domain-cluster in descending order.

Domain	Topics	Domain (cont.)	Topics (cont.)
Technology	16	Education	2
Informational	8	Energy	2
Geography	5	Military	2
Security	5	Health	1
Social	4	Political	1
Economic	3	Unidentified	1
Infrastructure	3		

The technology area consists of 16 topics. These topics are about algorithms, data, development, digitalisation, infrastructure, platforms, power issues, satellites, and transformation. It comes as no surprise that the technology area covers discussions on hardware, software, and telecommunications. This is essentially those fundamental technologies that have transformed societies from the Industrial Age to the Information Age. As computational power is increasing to process the vast amount of data that humans are creating, stored in data centres thanks to improved storage, transferred at speeds close to speed-of-light, with new advances on algorithms, machine-learning and "AI"-in general, humankind can really exploit the advantages these new technologies bring to improve human life. However, what is also evident is that the same technologies can be used for malicious purposes, but also to compete on how states should be governed (for example, democratic, autocratic), the world, different phenomena, and in essence, reality is perceived. Furthermore, these technologies are enabling ways to augment reality by visualising it, and interacting with it, in ways never possible before. The more of our senses that can be used to interact with these augmented realities, the more difficult it will be to discern "reality".

There are already challenges today where threat actors are conducting cyber-enabled disinformation operations targeting people and creating rifts in societies, but also rifts between societies and their elected. First, thanks to the ability to manipulate information in ways that it triggers people's worries', it is very easy to create these rifts, or to create doubt in an election system that the results are accurate, or if a vaccine is safe or not. As the ordinary citizen and user of these technologies can choose those who they want to interact with, it is very easy for them to choose the information that affirms/confirms those predetermined biases. Thus, threat actors are

directly and/or indirectly (through proxies) able to create informational "anti-access/area-denial spheres". Which leads to the Informational area.

The Informational area consists of 8 topics. The informational area is built on information that is visualised (e.g., maps), information created and shared by other humans, where they express their views on various topics, and related narratives. In essence, information within a given context framed from the perspectives of the beholder. Thanks to the technological areas above, humans are now able to exploit the benefits of new technological advancements, create content, tell stories, and reach audiences that pre-technological areas were not possible before. For example, in 2011, Pastor Terry Jones burned the Koran. Unconfirmed information states that the Afghanis were unaware of this case, until mainstream media began to report about it. Once the information ("news") reached Afghanistan, it led to massive violent protests where a mob attacked the UN camp in Mazar-e Sharif. Seven people were killed (Sief, 2011). What is maybe even somewhat ironic, is the title of the news article: "Florida pastor Terry Jones's Koran burning has far-reaching effect", which contains the words "far-reaching". Without these technologies, it is likely that the "Koran-burning"-information would have never reached the other side of the world, at the speed that it did, with the visual information that it did, resulting in the violent actions that it did.

The area of Geography consists of 5 topics. It consists of physical places (e.g., the Arctic and China), undersea places, and the role of geographical places in territorial disputes. The case for the arctic is quite self-explanatory. Climate change is having an impact on the arctic, melting the ice, leading to new opportunities to find resources and exploit those resources. This has led to a competition between various states, such as Canada, Russia and the U.S. China wants also to be a part of this and sees an opportunity. Therefore, geography, just like in the past, still can impact on world politics.

The Security area consists of 5 topics. It consists of lack of security, geopolitical risk, cybersecurity, attacks on democracy and water security. As noted, security, and the absence of security, is a broad topic. Cyber security, or the lack of it, can truly have an impact on world politics, and on international security, just like presented earlier. Therefore, cyber security, on all levels, and all domains, is important. The challenges are also many and range from identifying unintentional vulnerabilities in hardware and software, to the intentional search for vulnerabilities, to exploit them, and attack for espionage and/or criminal purposes. Other challenges that exist is public-private partnerships. The private sector owns much of the underlying infrastructure, and the private sector is responsible for the rapid development of technology and associated hardware and software. The public sector is not coping. Therefore, to ensure trust, certain measures need to be taken by both parties to ensure that sensitive information is shared securely and trust that the other party will not use it for their own advantages. In the U.K. for example, the National Cyber Security Centre (NCSC) has something known as "Industry 100". Industry 100 is "the principal initiative from the NCSC to facilitate close collaboration with the best and most diverse minds in UK industry" (NCSC, 2022). What is also noteworthy, is that "all i100 secondees will require a Security Check clearance" (NCSC, 2022).

The Security Check clearance is mandatory because it is likely that participants will have access to sensitive information which, if released publicly, could impact on British National Security. In the U.S., US Cyber Command has something known as "USCYBERCOM Industry Day 2022". This event gives USCYBERCOM the opportunity "to connect with private industry to communicate its capability requirements and challenges" (USCYBERCOM, 2022). Naturally, private industry partners are vetted before being accepted to the event.

What is the role of the private sector then? The private sector can also take their own initiatives. The Tech Accords, of which more than 100 companies have signed, has the mission to promote "a safer online world by fostering collaboration among global technology companies committed to protecting their customers and users and helping them defend against malicious threats" (Cybersecurity Tech Accord, 2022). Brad Smith, President & Vice Chair at Microsoft, argued the need for a "Digital Geneva Convention" (Smith, 2017). The problem that led to this initiative was naturally cybercriminal attacks, state, and state-sponsored cyber attacks. Smith (2017) argued that private sector companies, including Microsoft, are doing as much as they can, but Governments should now step up to. While the initiative is good, there are those who would see this as an opportunity to remove the existing Geneva Conventions.

Finally, looking at the Russo-Ukraine war, the private sector has shown its ability to contribute to cyber defence: identifying destructive malicious software, especially in the Ukrainian Rails System. What would the consequences have been if the malicious code had "detonated", rendering trains unable to roll, making it impossible for civilians to flee?

Acknowledgements

The views and opinions are of the author and do not necessarily reflect those of the GCSP. Also, I would like to thank Dr. Richard Vidgen for his advice on the CLR and STM and to contact Dr. Michael J. Mortenson, who in turn shared his expertise on the CLR and the STM.

References

Albert, M. (2022). Rudolf Kjellén. Accessed 21 June 2022. Retrieved from https://www.britannica.com/biography/Rudolf-Kjellen.

BBC. (2020). US charges Chinese Covid-19 research 'cyber-spies'. Accessed 22 June 2022. Retrieved from https://www.bbc.com/news/world-us-canada-53493028.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.

Cimpanu, C. (2020). Google said it took down ten influence operation campaigns in Q2 2020. Accessed 22 June 2022. Retrieved from https://www.zdnet.com/article/google-discloses-new-takedowns-of-influence-ops-on-its-sites/.

Cybersecurity Tech Accord. (2022). Mission Statement. Accessed 22 June 2022. Retrieved from https://cybertechaccord.org/about/.

Dean, G. (2021). Drivers face $3 gas prices after the Colonial Pipeline cyberattack, and some gas stations have run out completely. Accessed 21 June 2022. Retrieved from https://www.businessinsider.com/gas-prices-colonial-pipeline-cyberattack-fuel-east-coast-2021-5?r=US&IR=T.

Deudney, D.H. (2013). Geopolitics. Accessed 21 June 2022. Retrieved form https://www.britannica.com/topic/geopolitics.

Dunnett, O., Maclaren, A. S., Klinger, J., Lane, K. M. D., & Sage, D. (2019). Geographies of outer space: Progress and new opportunities. Progress in Human Geography, 43(2), 314–336. https://doi.org/10.1177/0309132517747727

Huskaj, G. (2019). The Current State of Research in Offensive Cyberspace Operations. In: Proceedings of the 18th European Conference on Cyber Warfare and Security, Academic Conferences and Publishing International Limited, 2019, p. 660-667.

Hyvärinen, N. (2017). APT1 - What happened next? Accessed 21 June 2022. Retrieved from https://blog.f-secure.com/apt1-what-happened-next/.

Kerner, S.M. (2022). Colonial Pipeline hack explained: Everything you need to know. Accessed 22 June 2022. Retrieved from https://www.techtarget.com/whatis/feature/Colonial-Pipeline-hack-explained-Everything-you-need-to-know.

Matt, C., Hess, T. & Benlian, A. (2015). Digital Transformation Strategies. Bus Inf Syst Eng 57, 339–343 (2015). https://doi.org/10.1007/s12599-015-0401-5.

Menn, J. & Bing, C. (2021). EXCLUSIVE Governments turn tables on ransomware gang REvil by pushing it offline. Accessed 22 June 2022. Retrieved from https://www.reuters.com/technology/exclusive-governments-turn-tables-ransomware-gang-revil-by-pushing-it-offline-2021-10-21/.

NCSC. (2020). Advisory: APT29 targets COVID-19 vaccine development. Accessed 22 June 2022. Retrieved from https://www.ncsc.gov.uk/news/advisory-apt29-targets-covid-19-vaccine-development.

NCSC. (2022). Industry 100. Accessed 22 June 2022. Retrieved from https://www.ncsc.gov.uk/section/industry-100/about.

Nikita, M. (2016). ldatuning. Retrieved from https://cran.r- project.org/web/packages/ldatuning/ldatuning.pdf

Online Etymology Dictionary. (n.d.). Digital (adj.). Online Etymology Dictionary. Accessed 21 June 2022. Retrieved from https://www.etymonline.com/word/digital#etymonline_v_31427.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation,

Sief, K. (2011). Florida pastor Terry Jones's Koran burning has far-reaching effect. Accessed 22 June 2022. Retrieved from https://www.washingtonpost.com/local/education/florida-pastor-terry-joness-koran-burning-has-far-reaching-effect/2011/04/02/AFpiFoQC_story.html.

Smith, B. (2017). The need for a Digital Geneva Convention. Accessed 22 June 2022. Retrieved from https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/.

Soldatkin, V. & Pamuk, H. (2021). Biden tells Putin certain cyberattacks should be 'off-limits'. Accessed 22 June 2022. Retrieved from https://www.reuters.com/technology/biden-tells-putin-certain-cyber-attacks-should-be-off-limits-2021-06-16/.

Turton, W. & Mehrotra, K. (2021). Hackers Breached Colonial Pipeline Using Compromised Password. Accessed 21 June 2022. Retrieved from https://www.bloomberg.com/news/articles/2021-06-04/hackers-breached-colonial-pipeline-using-compromised-password.

USCYBERCOM. (2022). USCYBERCOM Industry Day 2022. Accessed 22 June 2022. Retrieved from https://dreamport.tech/USCC-Industry-Day/.

Managing Variable Cyber Environments with Organizational Foresight and Resilience Thinking

Eveliina Hytönen, Jyri Rajamäki and Harri Ruoslahti
Laurea University of Applied Sciences, Espoo, Finland

eveliina.hytonen@laurea.fi
jyri.rajamaki@laurea.fi
harri.ruoslahti@laurea.fi

Abstract: Combining business continuity management (BCM) and systematic cyber threat intelligence (CTI) can improve cyber situational awareness to support decision-making through the phases of the resilience cycle (plan, absorb, recover, adapt) to ensure the continuity of organizational operations when encountered by cyber disruptions. End-user needs, human factors, high ethical standards, and social impacts can best be adapted when professionals from different fields work together with end-users to refine and co-develop selected tools into a platform. A resilience assessment that combines BCM and CTI enables 1) quick or detailed assessment of the investigated industry and its critical processes, 2) measurement of performance goals based on information received from end users, where artificial intelligence-based self-learning approaches can be used for functional descriptions, 3) information on the sensitivity of the investigated industry and vulnerability and 4) resilience and BCM throughout the entire resilience cycle. A new Horizon Europe project DYNAMO (Dynamic Resilience Assessment Method including a combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors) works towards combining BCM and CTI to generate a situational picture for decision support. Having this in mind, certain cybersecurity and BCM tools will be developed, refined, and integrated into the DYNAMO platform to provide decision support and awareness to chief information security officers, cybersecurity practitioners, and other stakeholders. This paper reports a case study that explores how combining CTI and BCM can help in the case of a cyber-attack. The research material consists of the news articles by the largest newspaper in Finland, Helsingin Sanomat (HS) of how the cyber attack against the therapy center Vastaamo progressed during the first week after the attack. The results show that cyber threat intelligence when flexibly integrated into the BCM approach could create better conditions for improved organizational foresight to react to unpredictable cyber threats to ensure business continuity.

Keywords: Cyber Threat Intelligence, Business Continuity Management, Resilience, Situational Awareness

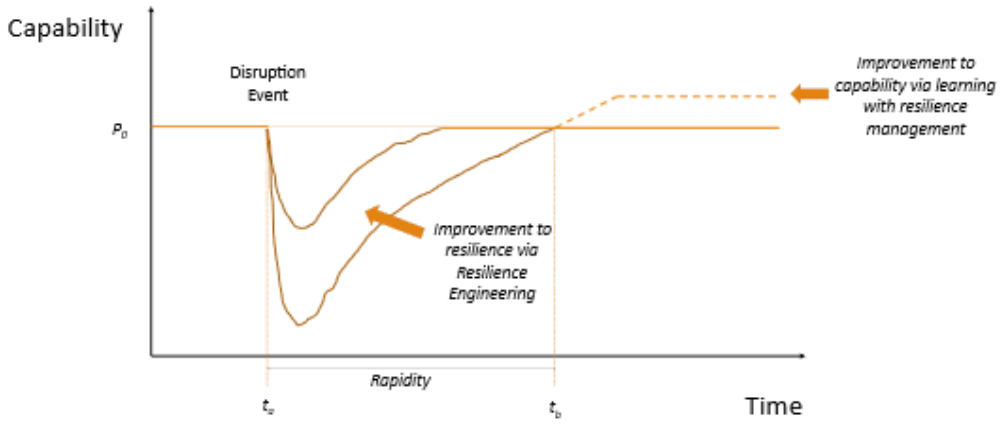
1. Introduction

Business continuity management (BCM), crisis management (CM), disaster recovery (DR), and resilience are related concepts, the purpose of which is to secure the critical functionality of the system in all situations. Risks and crises are often a derivative of external stressors, while the organization's resilience is more intrinsic, and in this sense, the priority of preventive behaviour at the organizational level is the preparation of various procedures to respond to crises or critical events (Linkov et al., 2014). Traditionally, BCM combines risk management and quality management. DR involves the re-establishment of systems and functions to recover from a disaster. The BC Plan lists the steps to be taken to ensure the continuity of critical business operations. (Sawalha, 2021.)

Figure 1 presents the management cycles of BCM and resilience management. The holistic BCM process identifies potential threatening impacts on the organization and provides a framework for developing resilience and the ability to respond effectively to protect the system and the interests of the key actors. The goal of resilience engineering is to improve resilience by reducing the drop in capability and speeding up recovery. The goal of resilience management is to learn from unwanted events and thus improve the system's capability.



a) Business continuity cycle



b) Resilience cycle and aspects of resilience engineering

Figure 1: Business continuity and resilience cycles

Resilience can be understood as business flexibility, endurance, or ability to recover from adverse events; an ability to recover or adapt to a new normal after a crisis or a critical event. The measure of resilience is the ability of an organization to minimize any negative impacts. Thus, resilient organizations can quickly adapt to wide systemic changes occurring in the organization, its value network, or impacting the entire society. Resilient organizations or networks have organizational stability, agility, and a culture that promotes situational awareness to detect and identify clues that may indicate the realization of risks for appropriate mitigation and reaction. (Palomäki, Roschier, Gilbert and Pokela, 2020.)

Examples of threats against critical infrastructures and vital societal services are numerous. The WannaCry attack in 2017 encrypted data and files on 230 000 computers in 150 countries and impaired the functionality of the National Health Service (NHS) in England (Ghafur et al., 2019), a cyber-attack halted the network of a Czech hospital in 2021 (Muthuppalaniappan and Stevenson, 2021), and in 2021 South Africa's port and rail operator, Transnet, experienced a cyberattack that caused a standstill in the movement of goods in some of its ports (South Africa Autos Report, 2021). A Finnish private psychotherapy service provider, Vastaamo, experienced cyber-attacks in 2018, 2019 and 2020, as the attacker stole thousands of records and then tried to blackmail individual patients directly threatening to expose documents containing everything from personal identity codes to therapy session transcripts (Tuttle, 2021). The financial losses forced Vastaamo to declare bankruptcy and close (Whitney, 2021).

The examples demonstrate vulnerabilities of critical services and how cyber threats may impact their business continuity. Systems that combine principles of business continuity with cyber threat warning systems can promote better preparedness and cyber resilience against cyber incidents. The goal of project DYNAMO is to combine cyber threat intelligence (CTI) and business continuity management (BCM) to generate a situational picture for decision support of critical sectors (DYNAMO, 2022). This study looks at the cyber-attack against the therapy center Vastaamo. The study adds in part to the practical body of knowledge that the DYNAMO project cumulates. The research question is: "*How to promote business continuity during cyber incidents*?".

2. Literature

Cyber security is still a young field of study, a consensus on the scientific definition of the term is lacking, but there is a common understanding that cyber security investigates how to secure cyberspace from threats and damage (Edgar and Manz, 2017). The Finnish security commission (Turvallisuuskomitea) defines cyber security as a state where cyberspace can be trusted, and its functions become secured (Turvallisuuskomitea, 2018).

The National Initiative for Cybersecurity Careers and Studies (NICCS) defines cybersecurity as "[t]he activity or process, ability or capability, or state whereby information and communications systems and the information contained therein are protected from and/or defended against damage, unauthorized use or modification, or exploitation" (NICCS, 2017). Dalal et al. (2021, p. 4) include the human-factor in the definition of cyber security, more specifically the definition of organizational cyber security as "the efforts organizations take to protect and

defend their information assets, regardless of the form in which those assets exist, from threats internal and external to the organization".

Cyberspace can be defined in many ways. This study understands it as an entity that combines three viewpoints in the field of research: information, technology, and social viewpoint as presented in the Figure 2. The healthcare system with its various information systems is an example of cyberspace (Lehto, Pöyhönen and Lehto, 2019).

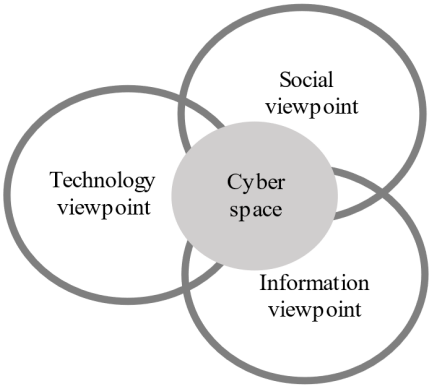


Figure 2: The dimensions of cyberspace (modified Edgar and Manz, 2017)

The basis of business continuity management is to combine risk and quality management principles based on probability-based quantitative methodology, which is useful in calculating predictable or calculable stress situations. However, securing systems with risk and business management-based solutions may be unrealistic for many systems for two reasons: (1) current interconnected social, technical, and economical networks form such complex systems that considering all possible risks may be very time-consuming and costly; (2) they may consider only events that are known or have happened before. (Linkov et al., 2014.)

The concept of resilience is based on two very different scientific traditions: (1) psychology, where resilience is referred to as affecting how an individual recovers from personal trauma or crisis; (2) Ecological systems thinking that looks at resilience as a nonlinear dynamics of recovery or adaption. Later resilience has been used in studies looking at ecosystems management, sustainable development, climate change, and politics. (Linkov and Kott, 2018.)

National Academy of Sciences (2012) define resilience as "the ability to prepare and plan for, absorb, recover from, or more successfully adapt to actual or potential adverse events". On the organizational level, the term 'resilience' can describe the inherent characteristics of the organizations that can respond more quickly, recover faster, or develop more unusual ways of doing business under pressure than others (Vogus and Sutcliffe, 2007). Resilience can also be seen as the capability of a system to recover in the middle of shocks or stressful events over time (Palma-Oliveira and Trump, 2016). In a system (e.g., an organization) there are multiple different levels that are interrelated. This implies interaction between those levels that can be called factors and sub-systems (Palma-Oliveira and Trump, 2016) or domains (Linkov et al., 2013). Herbane (2016) states that in the field of economics, resilience is often characterized as the aspiration or aim of business continuity management systems, or as the articulation of the future state of organization following achievement of its strategy. Business Continuity Management is one of several activities to support resilience.

A systems approach to resilience can be described and analysed using the Resilience Matrix, that maps system domains across an event management cycle of resilience functions (Linkov, 2013). The event management cycle or the resilience cycle can be simplified into four functions: plan, absorb, recover, and adapt. Decision-making can benefit from using the phases of resilience by developing e.g., measurement methods and practices, which can be used to assess the threats, probabilities, and impacts on the critical functions of the organizations (Juntunen, 2014).

Cyber threat intelligence (CTI) is actionable threat information that is relevant to a certain organization and that requires careful attention. CTI consists of activities such as acquiring, processing, and analyzing threat information to identify, track, and predict threats, risks, and opportunities inside the cyber domain, and thus to engineer more precise defense strategies. CTI can have a crucial role in directing organisational behaviour in

prevention, detection, and response to cyber-attacks. It can enable better decision making at different levels in the organization, including the strategic, tactical, and operational levels. (Shin and Lowry, 2020.)

Kotsias et al. (2022) state that even though firms are fully compliant with industry "best practice" cyber security standards, they may not be able to cope with well organised, sophisticated cyber-attacks. Leveraging cyber-threat intelligence (CTI) to direct cyber-defence is key in resolving the asymmetry. The adoption of CTI can pose many challenges to companies. There might be limited knowhow about the deployment and use of CTI. Moreover, many organizations would need to transform their cybersecurity practice from a reactive logic towards a proactive logic driven by CTI.

3. Methodology

Case study research aims at providing deep and detailed information on the subject of the case study to support development activities. Development work is based on prior theories, knowledge, and study that deal with similar problems as the study at hand. Case study often aims at answering the questions 'how?' and 'why?'. The case study takes into account the connections and social situations in time and space. A case study is a well-suited approach when trying to deeply understand the situation of the case organization by studying it in its own environment to produce a solution to a given problem or provide researched development suggestions (Yin, 2009).

This case study builds on the news articles by the largest newspaper in Finland, Helsingin Sanomat (HS) of how the cyber attack against the therapy center Vastaamo progressed during the first week. The material is a sample of 16 HS articles between 24.10. – 29.10.2020. To answer the research question, relevant data were extracted into a table that was specifically developed for this study.

4. Case cyber-attack against Vastaamo

The results of this case study are based on 16 publicized newspaper articles on how the attack progressed during the first news week (see Figure 3). These articles deal with how the cyber-attack and ransom case, the ensuing police investigation, and the reactions and situation of the victims evolved. This part briefly presents the sample articles.

Table 1: Articles published in Helsingin Sanomat between 24 October and 29 October in 2020

Articles published in Helsingin Sanomat between 24 October and 29 October in 2020			
Article	**Date**	**Title (translations by the authors)**	**Title in Finnish**
1	24.10.	Cyber-attack against Vastaamo got hackers chasing the thief	Vastaamon tietomurto sai hakkerit jahtaamaan varasta
2	24.10.	This feels extremely hurtful	Tuntuu äärimmäisen loukkaavalta
3	25.10.	Cowardly to blackmail with hacked patient information	Hakkeroiduilla potilastiedoilla kiristäminen on raukkamaista
4	25.10.	Customers blackmailed directly	Asiakkaita kiristetty suoraan
5	26.10.	Data breach reported weeks after notifying police	Murrosta tiedotettiin viikkoja rikosilmoituksen jälkeen
6	26.10.	Experts: data breach victims in the thousands	Asiantuntijat: Tietomurron uhreja on jopa tuhansia
7	26.10.	Strong feelings at this moment are insecurity and disbelief	Vahvoja tunteita tällä hetkellä ovat turvattomuus ja epäusko
8	27.10.	Psychotherapy centre Vastaamo has a bad reputation in its field	Psykoterapiakeskus Vastaamolla on alalla jopa surkea maine
9	27.10.	Data breach did not surprise healthcare actors	Tietomurto ei yllättänyt sote-alan toimijoita
10	27.10.	Forensic psychologist: exceptionally shrewd case	Oikeuspsykiatri: Tapaus on poikkeuksellisen häikäilemätön
11	28.10.	Damages may take years	Vahingonkorvauksia voi joutua odottamaan vuosia

Articles published in Helsingin Sanomat between 24 October and 29 October in 2020			
Article	**Date**	**Title (translations by the authors)**	**Title in Finnish**
12	28.10.	Psychotherapist: important now to support victims	Psykoterapeutti: Uhrien tukeminen on nyt tärkeää
13	28.10.	Ideas on how to help the victims	Keinoja tietomurron uhrien avuksi ideoidaan
14	29.10.	Vastaamo may be the largest crime skein in Finland	Vastaamo voi olla isoin rikosvyyhti Suomessa
15	29.10.	F-Secure trying to track Vastaamo blackmailer	F-Secure yrittää jäljittää Vastaamo-kiristäjää
16	29.10.	Experts: data breach easy	Asiantuntijat: Tietomurto oli helppo

Article 1 HS 24.10.2020 C*yber-attack against Vastaamo got hackers chasing the thief*. Due to the cyber-attack against the psychotherapy center Vastaamo, up to tens of thousands of patients' information is suspected to have ended in the hands of a hostile cyber attacker. Vastaamo confirms the cyber-attack, and that the attacker is demanding 450.000 Euro as bitcoins, or the information of 100 people will be made public on the Internet every day. The police are investigating this as a cyber-attack and a felony of distributing private information. The cyber experts interviewed suspected that the information had perhaps been only secured by a default password and thus, had been quite easily accessible. In their view the Vastaamo network had been neglected (e.g., many network servers had not been updated in years) and network addresses pointing to the patient records or Vastaamo Intranet had been publicly visible. The attacker has notified the attack in a discussion.

Article 2 HS 24.10.2020 *This feels extremely hurtful.* The psychotherapy center Vastaamo informed its customers about the cyber-attack and ransom request first only on its website. According to the deputy data protection ombudsman this was not effective enough, and that Vastaamo had been ordered to personally inform everyone in their registers. The information of 300 people has been made public in the Tor network. A victim of the cyber-attack, who was interviewed by HS says that he is worried about patient safety and that he is hesitant to see a therapist in the future.

Article 3 25.10. *Cowardly to blackmail with hacked patient information*. The hacker removed on Friday (23.10.) the sensitive patient information from the Tor network, and them reloaded them. Apparently, the patient data was insufficiently encrypted, as they could be hacked so easily. This makes it difficult for patients to believe in the privacy of treatment. There is only a minor probability that the attackers are caught. They would have to make a mistake to be caught, says an expert from F-Secure.

Article 4 25.10. *Customers blackmailed directly*. Several victims of the data breach have reported that on Saturday (24.10.) at around 19:30 they have received a direct ransom message demanding them to pay Bitcoins in the sum of 200 Euro (if paid within 24 hours) or 500 Euro (if paid within 48 hours). The ransom message states that the victims must pay, because Vastaamo refuses to take responsibility and pay.

Article 5 HS 26.10. Data breach reported weeks after notifying police. The police notes on Sunday (25.10.) that they have so far received thousands of reports of the data breach crime. The police says that they do not know the nationality or the location of the attacker, and if the ransom messages are sent by the attacker. Vastaamo says, they had reported the crime and notified the Cyber security center, The National Supervisory Authority for Welfare and Health, and the data protection ombudsman as soon as they gained information of what had happened. Vastaamo reveals that customer information may have disappeared already in data breaches in 2018 and 2019. The Ministry of the interior took stand, by informing on Saturday (24.10) that the victims of the data breach need urgent help.

Article 6 26.10. *Experts: data breach victims in the thousands*. The blackmailer shared for a while a folder with personal information from up to 2.000 patients in the dark web. The leaked information contains patient names, personal identity numbers, contact information, notes from therapy sessions and medical reports.

Article 7 26.10. *Strong feelings at this moment are insecurity and disbelief*. One victim says that it was difficult to find information about what to do. There were no clear instructions before this day. The first days were filled with uncertainty and silence. He has suffered from insomnia and reports seeing a great worry of those who don't have the resources to do anything about this. People are totally devastated; small regrets will not be enough, something bigger must be done.

Article 8 27.10. *Psychotherapy center Vastaamo has a bad reputation in its field*. The CEO of Vastaamo was let go because of being suspected of keeping the data breach of the company server a secret for over a year. The ex-CEO claims that the November 2018 breach only became evident when there was an investigation in October

2020. The National Supervisory Authority for Welfare and Health notes that the Vastaamo case has raised the question of a need to strengthen cyber security and the level of its control.

Article 9 27.10. *Data breach did not surprise healthcare actors*. An expert from the Association of Finnish Local and Regional Authorities notes that the private operators of the healthcare sector have a lot to develop as cyber security is concerned. The approximately 18.000 companies are of different sizes and may have very different levels of investments in cyber security.

Article 10 27.10. *Forensic psychologist: exceptionally shrewd case*. The Head forensic psychologist estimates that the Vastaamo data breach and the ensuing ransom demands are exceptionally shrewd and show an exceptionally far-taken indifference toward weaker people.

Article 11 28.10. *Damages may take years*. Responsibility to pay the damages is with people responsible for the data breach, with the attacker (if they are ever caught), with Vastaamo (if it has not complied with data protection regulations), and with those healthcare districts that have bought Vastaamo services. This is a case of distributing information that violates private life, notes a professor of public law. The first step is to determine if the register-holder has acted according to data protection regulations.

Article 12 HS 28.10. *Psychoterapist: important now to support victims*. The feeling of safety may be become severely disturbed, and it is unrealistic to expect that they all can report the crime, comments a Head of a psychotherapy center. Data breach experts have told that they have seen the information from as many as 2.000 patients.

Article 13 28.10. *Ideas on how to help the victims*. The Minister of Justice notes that one way to help the Vastaamo victims could be to provide a rapid change of personal identification numbers. The Ministry of Finance is preparing legislation for expediated change of personal identification numbers.

Article 14 29.10. *Vastaamo may be the largest crime skein in Finland*. The National Bureau of Investigation has received over 15.000 reports of crime regarding the Vastaamo data breach and ransom case. The deadlines given by the blackmailer have passed, but the police have no information of the blackmailer taking any new action. The Minister of science and culture has commented that similar data breaches to Vastaamo should never happen again. This data breach has been compared to a major disaster. Customer information regulations have to be changed so that similar companies will have tighter cyber security requirements.

Article 15 29.10. *F-Secure trying to track Vastaamo blackmailer*. The cybersecurity company F-Secure is asking Vastaamo victims to contact their research director to collect information on the Bitcoin wallets to try to track where the ransom money has gone. Bitcoin is based on blockchains with unchanged public records of transactions that F-Secure will try to follow.

Article 16 29.10. *Experts: data breach easy*. Cybersecurity experts interviewed note that the Vastaamo network had been configured in a way that the data breach could be easily done. When googling for information on Vastaamo, a link to Vastaamo's public materials and references to Vastaamo's own instructions under the name 'Patient register' were easily available. Someone claiming to be the attacker revealed in the dark web that the patient register was obtained by using a default user and passwords. An expert interviewed by the HS found some twenty vulnerabilities in the Vastaamo server programs that pertained to the web address of the patient register.

5. Results

Results and conclusions are based on the data collected from the 16 sample articles. The results are discussed in the framework of a resilience cycle with six phases: Prepare, Prevent, Protect, Response, Recover, and Learn/adapt (Figure 4).

Table 2: Results of the study in the framework of resilience cycle

	Prepare	Prevent	Protect	Response	Recover	Learn/adapt
Cyber Threat Intelligence	Customer information disappeared already 2018 and 2019	Vulnerability of today's information society		Nationality of attacker nnot know Notify ransom demand to police	Vastaamo ex-CEO claims 2018 breach became first evident when in October 2020	Private healthcare operators have a lot to develop as cyber security is concerned

	Prepare	Prevent	Protect	Response	Recover	Learn/adapt
Identifying risks	Vastaamo network had been configured in a way that the data breach could be easily done	Visible patient records / Intranet Patient data insufficiently encrypted	Only default password in use 'Patient register' = may have attracted the attacker	How to prevent blackmail of individual patients or to begin using personal data to commit identity thefts	Information from as many as 2.000 patients visible (28.10.) Over 15.000 data breach and ransom case (29.10.)	Patient safety / readiness to see a therapist in the future Belief in privacy of treatment
Critical activities	Some twenty vulnerabilities in server programs to patient register web address		Network neglected	Vastaamo refuses to pay	Ministry of the interior: victims of data breach need urgent help	Determine if acted according to data protection regulations
Key personnel				Police be given peace to work	CEO let go / suspected of keeping data breach secret for over a year	Cybersecurity company track where the Bitcoin ransom gone
Guidelines and procedures			Raised need to strengthen cyber security and the level of its control	Paying ransom will not get information back.	Victims need a rapid change of personal identification numbers	Similar companies to Vastaamo will have tighter cyber security requirements
Open communication			Victims informed first only on Vastaamo website (24.10.) Data breach reported only weeks after notifying police	Direct ransom demands in Bitcoins Victims' first days filled with uncertainty and silence	Clear instructions have only been available during the last day (26.10.)	Data breach ransom demands exceptionally shrewd and showing indifference toward weaker people

Cyber Threat Intelligence seems to have been neglected in the company. Customer information apparently disappeared already 2018 and 2019. This demonstrates the potential vulnerability of personal data today's information society, and the need for developing cyber security of private healthcare operators. With the help of CTI it would be possible to gain information about possible threats and support the situational awareness and decision making in different stages of resilience.

It seems that identifying risks was also neglected. The Vastaamo network had been configured in a way that the data breach could be easily done, and e.g., the location of patient records and Intranet had been left visible for those who know how to look for them. The title 'Patient register' may have attracted the attacker. Patient data was insufficiently encrypted, and only default passwords in use.

Critical activities overlooked include some twenty vulnerabilities in server programs. The victims needed urgent help which may involve the wider society, and from a legal standpoint it must be determined if the company acted according to data protection regulations. Guidelines and procedures were not enough to ensure the protection of the data. This raises a clear need to strengthen cyber security and the level of control for similar companies.

Communication was not open. The victims were first only informed on the Vastaamo website, and this happened weeks after the police had been notified of the data breach. Once the company refused to pay, the victims began to receive direct ransom demands in Bitcoins. The victims didn't get clear instructions until after a few days. This data breach was deemed exceptionally shrewd as it showed a grave indifference toward weaker people.

6. Conclusions

Working in critical sectors is the backbone of society. Strong interconnections highlight vulnerabilities and make it possible for the effect to cascade between organizations. More and more cyber threats target critical infrastructures, threatening their business continuity. When assessing cyber resilience, resilience phases and business continuity management should be combined with cybersecurity processes and principles.

Understanding how functions change in the different stages of resilience can help to quickly detect threats to appropriately counter them for added endurance and stability.

The case of Vastaamo is one indication of how the know-how and professionalism of many companies in the SME sector are not sufficient to develop information security and maintain the continuity of their operations. Data breaches can test a company's ability to continue its operations, and especially sectors critical to society, such as health care, directly affect society's well-being.

Traditional risk assessments and information security management processes do not consider complex interdependencies and their effects on the continuity of operations. By understanding the different stages of resilience, methods can be developed to assess threats, as well as their probabilities and effects on operations. By combining the principles of business continuity with cyber situational awareness created through CTI, the company's resilience can be increased.

Procedures and tools that improve the resilience and continuity of information security could include activities that identify the sector's critical information, functions, and reserves, securely store critical protected information, test the effectiveness of contingency plans and train personnel and stakeholders. Combining the principles of business continuity management and systematic cyber threat assessment can be a way to increase cyber situational awareness, which is needed to support decision-making in all different phases of resilience.

Important topics for further research are the deepening of knowledge and understanding regarding the methods and tools that improve the resilience and continuity of information security mentioned above. Considering the phases of resilience as a reference framework can increase systematicity and the further research and practical development of methods and tools that improve the resilience and continuity of information security.

Acknowledgements

Acknowledgement is paid to DYNAMO Project, funded by the European Union under grant agreement no. 101069601. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

Dalal, R.S. *et al.* (2022) 'Organizational science and cybersecurity: abundant opportunities for research at the interface', *Journal of Business & Psychology*, 37(1), pp. 1–29. doi:10.1007/s10869-021-09732-9.

DYNAMO (2022) 'Dynamic Resilience Assessment Method including combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors', Webpage. Available: https://horizon-dynamo.eu/

Edgar, T. and Manz, D. (2017) *Research Methods for Cyber Security*, Cambridge, MA: Elsevier.

Ghafur, S., Grass, E., Jennings, N.R. and Darzi, A., 2019. 'The challenges of cybersecurity in health care: the UK National Health Service as a case study', *The Lancet Digital Health*, *1*(1), pp. e10-e12.

Herbane, B. (2016) *A Business Continuity Perspective on Organisational Resilience, in Resource Guide on Resilience*, EPFL International Risk Governance Center, Lausanne, available at https://www.irgc.org/irgc-resourceguide-on-resilience/

Juntunen, T. (2014) 'Kohti varautumisen ja selviytymisen kulttuuria', *Kriittisiä näkökulmia resilienssiin,* SPEK puheenvuoroja 2, 6-7.

Kotsias, J., Ahmad, A. and Scheepers, R. (2022) 'Adopting and integrating cyber-threat intelligence in a commercial organisation', *European Journal of Information Systems*, pp. 1–17. doi:10.1080/0960085x.2022.2088414.

Lehto, M.; Pöyhönen, J. and Lehto, M. (2019) *Kyberturvallisuus sosiaali- ja terveydenhuollossa, Loppuraportti, Vol. 2,* Jyväskylä: Jyväskylän yliopiston IT-tiedekunta.

Linkov, I.; Bridges, T.; Creutzig, F.; Decker, J.; Fox-Lent, C.; Kröger, W.;. . . and Thiel-Clemen, T. (2014) 'Changing the resilience paradigm', *Nature Climate Change*, 4, 407-409.

Linkov, I. and Kott, A. (2018) *Fundamental Concepts of Cyber Resilience: Introduction and Overview*, Cornell University Library, arXiv.org, Ithaca.

Muthuppalaniappan, M. and Stevenson, K. (2021) 'Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health', *International journal for quality in health care : journal of the International Society for Quality in Health Care*, 33(1). doi:10.1093/intqhc/mzaa117.

Munk, S. (2018) 'Interoperability Services Supporting Information Exchange Between Cybersecurity Organisations1'. *AARMS, 17*(3), 131-148.

National Academies of Sciences, Engineering, and Medicine. (2012) *Disaster Resilience: A National Imperative. Washington, DC*, The National Academies Press. https://doi.org/10.17226/13457.

Palma-Oliveira, J.M. and Trump, B.D. (2016) *Modern resilience: Moving without movement, in Resource Guide on Resilience*, EPFL International Risk Governance Center, Lausanne, available at https://www.irgc.org/irgc-resourceguide-on-resilience/

Palomäki, S.; Roschier, S.; Gilbert, Y. and Pokela, P. (2020) *Kestävän tuotannon resilienssi: Kuinka varautua kriiseihin ja kasvaa kestävästi,* Helsinki: Business Finland.

Sedenberg, E. M. and Dempsey, J. X. (31. May 2018) *Cybersecurity Information Sharing Governance Structures: An Ecosystem of Diversity, Trust, and Tradeoffs*, Retrieved 30.3.2019 from https://arxiv.org/abs/1805.12266

Shin, B., and Lowry, P. B. (2020) 'A review and theoretical explanation of the 'cyberthreat-intelligence (cti) capability' that needs to be fostered in information security practitioners and how this can be accomplished', *Computers & Security*, 92, 101761. https://doi.org/10. 1016/j.cose.2020.101761

South Africa Autos Report (2021) *Fitch Solutions Country Industry Reports,* p. 1., Q4 2021.

Turvallisuuskomitea. (2018) *Kyberturvallisuuden sanasto*. Helsinki: Sanastokeskus TSK ry.

Tuttle, H. (2021) 'Ransomware Attackers Turn to Double Extortion', *Risk Management*, vol. 68, no. 2, 2021, pp. 8-9.

Vogus, T.J. and Sutcliffe, K.M. (2007) 'Organizational resilience: towards a theory and research agenda', Paper presented at the *IEEE International Conference on Systems, Man and Cybernetics*.

Whitney, L. (2021) 'Ransomware attackers are now using triple extortion tactics', *Tech Republic*, May 12, 2021.https://www.techrepublic.com/article/ransomware-attackers-are-now-using-triple-extortion-tactics/

Yin, R. (2009). *Case study research: Design and methods* (Ed. 4), Thousand Oaks: Sage.

Fingerprinting Network Sessions for the Discovery of Cyber Threats

C.D. Klopper[1], J.H.P. Eloff[2]
[1]Masters in IT Big Data Science, University of Pretoria, Pretoria, South Africa
[2]Department Computer Science, University of Pretoria, Pretoria, South Africa

u26305667@tuks.co.za
jan.eloff@up.ac.za

Abstract: Artificial intelligence (AI) assisted cyber-attacks, within the network cybersecurity domain, have evolved to be more successful at every phase of the cyber threat lifecycle. This involves, amongst other tasks, reconnaissance, weaponisation, delivery, exploitation, installation, command & control, and actions. The result has been AI-enhanced attacks, such as DeepLocker, self-learning malware and MalGan, which are highly targeted and undetectable, and automatically exploit vulnerabilities in existing cyber defence systems. Countermeasures would require significant improvements in the efficacy of existing cyber defence systems to enable the discovery and detection of AI-enhanced attacks in networks in general. The challenge is that rule-and-anomaly-based intrusion detection approaches would need to be evolved into a dynamic self-learning approach before being able to discover "undetectable" network threats. The problem is that, when considering current state-of-the-art network cybersecurity countermeasures, this has not yet been achieved. One of the key challenges in achieving this is the inability to extract meaningful information from network packets. The novel solution proposed in this paper is to fingerprint network sessions. Each fingerprint is represented by a two-dimensional matrix that can be visualised, comprising a unique session key, the protocol discourse and the transmitted data. This is achieved by extracting information, summarising network session key events, encoding the received data, and merging it with existing fingerprints. The unique key and transmitted data are encoded using a Hilbert curve, while the protocol discourse is encoded into a tornado diagram. The resulting visualised network session fingerprints reveal hidden patterns that are ideal for subsequent pattern recognition, reinforcement learning (RL) or support vector machines (SVM) training to discover AI-enhanced cyber threats as they evolve.

Keywords: Cybersecurity, Cyber Defence, Network Session Fingerprints, Packet-Based Feature Extraction, Visual Data Mining

1. Introduction

One of the biggest challenges of the 21st century (Sobers 2022) is defending a company's cyber assets from cyber-attacks. According to Sobers (2022), there has been a significant increase in hacking and data breaches since the start of COVID-19. Findings in (Sobers 2022) characterise unprotected data and poor cyber security practices as one of the key challenges in improving cyber defence.

A widely used term that describes the importance of cyber defence is "zero-day attack" (Kaspersky 2022). Zero-day attack refers to a situation where a large organisation or an anti-virus company has discovered a vulnerability for which there are zero days to mitigate. With AI-assisted cyber-attacks within the cybersecurity domain, "zero-day attacks" will be extended to the time when new advancements in new cyber threats are discovered. The most notable example is ransomware. According to Indusface (2022), in 2021 ransomware attacks increased by 92.7%. This is likely due to the increased use of more advanced machine learning (ML) and deep learning (DL) techniques that make new ransomware malware more undetectable (Du, Raza, Ahmad, Alam, Dar and Habib 2022).

Cybercriminals using AI, BotNets and generative adversarial networks (GAN) have advanced their ability to perform highly sophisticated AI-assisted attacks that have become increasingly more difficult to discover, detect and protect against (Kaloudi and Li 2020) and (Nguyen and Reddi 2019).

The University of New South Wales (UNSW) and the Australian Centre for Cyber Security (Moustafa and Slay 2015) performed a simulation using an IXIA PerfectStorm tool to record data for research in cyber security. UNSW performed intrusion detection during the simulation on the nine types of attacks that were synthesised during the simulation. The intrusion detection and prevention performance are depicted in Figure 1.

From Figure 1, four cyber-attacks were completely undetected (analysis, backdoors, shellcode and worms) including "zero-day attacks". Only a fraction of the attacks were detected, namely denial of service (DoS - 3.7%), exploits (0.6%), fuzzers (4.2%), generic (3.5%) and reconnaissance attacks (0.2%). This result proves the inability of traditional rule and anomaly threat detection as a cyber defence to successfully detect threats that were simulated. State-of-the-art cyber defence systems should be capable of discovering and detecting 99% of threats for all categories to tolerate automated responses.

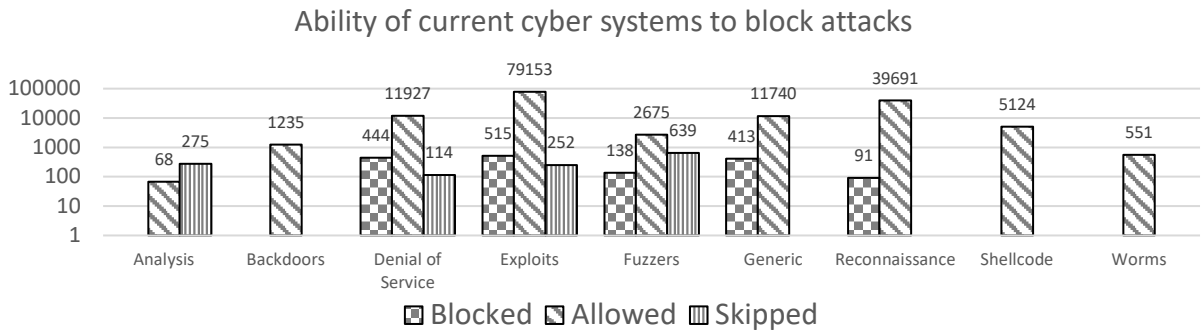


Figure 1: UNSW cyber simulation results

The simulation performed by UNSW was completed in 2015; since then, the cyber threat landscape has evolved with AI-assisted cyber-attacks. In a study conducted by Kaloudi and Li (2020), AI-assisted attacks were identified in nearly every phase of the cyber threat lifecycle.

The cyber threat lifecycle stipulates the general tasks a cyber attacker needs to complete to infiltrate an organisation. The key tasks are:

- During the reconnaissance task, a cyber attacker collects information and intelligence to aid the planned cyber-attack.
- During the weaponization task, the collected information is used to improve the effectiveness of the cyber-attack.
- During the delivery task, the cyber-attack needs to successfully bypass the boundary safeguards such as intrusion detection defence systems.
- During the exploitation task, the cyber-attack attempts to infiltrate the target organisation's network.
- During the installation task, the cyber-attack obtains a secure connection to open the organisation’s network for malicious attacks.
- During the command & control task, the cyber-attacker has remote control of the organisation’s network.
- During the actions task, the cyber-attacker executes the intended malicious activity.

In Kaloudi and Li (2020), the following key cyber-threats were identified that demonstrated the highest levels of AI-assistance (Refer to Kaloudi and Li (2020) for a complete list of threats across the cyber threat lifecycle):

- Smart Malware is a self-learning malware that can deploy malicious attacks, while it seems like accidental failures by manipulating control systems. Smart Malware represents advancements in AI-targeted reconnaissance.
- MalGAN is an algorithm based on a GAN that generates adversarial malware to bypass black-box malware detection systems. MalGAN can, upon detection, generate new malware examples to remain undetectable. DeepDGA is a next-generation approach to bypass domain name detection systems based on deep neural networks (DNN). Both MalGan and DeepDGA represent advancements in AI-concealed delivery.
- DeepLocker is an extremely targeted evasive malware and only deploys once it confirms its target. No defences have been implemented against DeepLocker. DeepLocker represents advances in AI-multilayered C2.

It is concluded in (Kaloudi and Li 2020) and (Nguyen and Reddi 2019) that AI-enhanced attacks have evolved to highly targeted, undetectable malware that automatically exploits vulnerabilities in existing cyber defence systems.

Therefore, with cyber threats evolving with AI-assistance across every phase of the cyber threat lifecycle (Kaloudi and Li 2020), cyber defence systems require significant improvements to improve efficacy in networks in general. The challenge is that rule and anomaly based intrusion detection approaches would need to be evolved into a dynamic self-learning approach (Kaloudi and Li 2020) which is capable of reacting to dynamic changing threats in real time.

According to Kaloudi and Li (2020), a cyber defence system capable of defending against AI-assisted cyber threats needs to incorporate RL methods. This is because a system based on RL methods can decipher complex and dynamic changes in cyber threats by learning from its experiences, and then respond to the threat in real-time.

The problem is that when considering the current state-of-the-art network cybersecurity countermeasures, a dynamic RL-based cyber defence has not yet been achieved. Caminero, Lopez-Martin, and Carro (2019) developed an adversarial environment (AE) RL algorithm for intrusion detection. In Caminero et al. (2019) and Lopez-Martin, Carro and Sanchez-Esguevillas (2020), the input source was pre-determined network features, which are not applicable to a real-time environment. The novel AE-RL model proposed by Caminero et al. (2019) scored a precision of 81% (DoS), 49% (Probe), 79% (R2L) and 9% (U2R), which is not suitable for an automated response. Lopez-Martin et al. (2020) expanded to additional RL models, including deep q network (DQN), deep dueling q network (DDQN), policy gradient and actor critic techniques. The DDQN scored a precision of 92% (flooding), 0% (impersonation), 93% (injection) and 96% (benign), which is not suitable for an automated response. Malialis (2014) developed a distributed RL for network intrusion response in real-time. In Malialis (2014), the input source was real-time network packets. The proposal is to use a distributed RL defence system to throttle distributed denial of service (DDoS) attacks. The group score (considering various input sources) for % of legitimate traffic unaffected was 72%.

One of the key challenges in achieving a dynamic RL-based cyber defence is the inability to extract meaningful information from raw network packets, which is crucial for an RL-based cyber defence to self-learn. This is because, once in operation, a dynamic RL self-learning cyber defence will need to detect threats in real time from network traffic. To overcome this challenge, the novel solution proposed in this paper is to fingerprint network sessions, comprised of the relevant raw network packets. Raw network packets contain a lot of information and vary in size. A fingerprint will overcome this by extracting only the meaningful information, while standardising the size of the input into the RL model. This allows the RL cyber defence system to learn from raw network packets, without being encumbered by irrelevant information. Each fingerprint is represented by a two-dimensional visualisation and comprises a unique session key that is available in network feature datasets, and the protocol discourse and transmitted data which can be extracted only from raw network packets.

Therefore, this is a novel topic for research, since no previous research considered RL techniques using raw network packets, nor using raw network packets to determine and encode the protocol discourse and transmitted data as proposed in this paper. The main research question for the research project at hand is as follows: “Can a fingerprint be developed that captures the hidden patterns contained in network sessions, which were previously undetectable, and can it be visually encoded?”

2. Relevant and Background Work

There are two different areas of research that contribute to the research at hand, namely:

1. Research that focuses on extracting meaningful features from network traces; and
2. Research that focuses on visual data mining and encoding data into visual information.

The relevant research for each field of research will be discussed in more detail in the remaining sections.

2.1 Extracting features from network traces in the form of PCAP files

PCAP files are generated from performing network traces. The contents of the PCAP are all the network files that were captured during a network trace, which include source and destination IP addresses, ports, protocol data and the actual data to be transmitted.

Several papers have developed algorithms that determine and process features for training (ML) directly from PCAP files. To evaluate these approaches, each will be considered in the method by which they aggregated the network traffic (Zhou, Niu, Zhang, Peng, Wu, and Hu 2020). in Table 1.

Table 1: Summary of relevant research in extracting features from network traces

Aggregation	Description	Relevant research	Performance
▪ Protocol level	▪ All network packets transmitted are considered per protocol	▪ Feature Map (Sun, Ochiai, and Esaki 2020) – uses a Hilbert curve to develop an activity map for 9 protocols for every 128 seconds and federated learning	▪ The highest accuracy scores were 92% and 87%.

Aggregation	Description	Relevant research	Performance
▪ Connection level	▪ All individual network traffic transmitted are considered per packet	▪ Packet2Vec (Goodman, Zimmerman, and Hudson 2020) – uses a modified embedding technique (Word2Vec) to vectorise network packets for automated feature extraction	▪ 0.46% malware samples: ▪ 63% precision, 93% recall, 75.1% F1
▪ Network flow level	▪ All network packets transmitted are considered per: ▪ two hosts, with specific IP addresses, ports, and protocol ▪ for each flow separately (there could be multiple flows between two hosts, IPs and ports that represent distinct communication events)	▪ Graph model (Nari and Ghorbani 2013) – uses dependencies between flows and flow information to develop behavioural profiles	▪ Existing anti-virus more superior
		▪ PcapNg (Velea, Ciobanu, Gurzau, and Patriciu 2017) – is a proposal to extract and store key network flow information (not applied to intrusion detection)	▪ Untested approach
		▪ Wavelet analysis (Du, Ma, Li, Li, Sun, and Liu 2018) – transforms the network flow into Daubechies1, Coifiets1 and Discrete Meyer wave functions, and then extracts wave features for anomaly detection	▪ 89% test precision (unclear amount of malware samples)
▪ Session level	▪ All network packets transmitted are considered per: ▪ two hosts, with specific IP addresses, ports and protocol	▪ DL-IDS (Sun et al. 2020) – uses CNN and LSTM to extract spatial and temporal features (developed a transmitted data image with one-hot-encoding)	▪ Overall F1 score of 93.32%, with a low sample % of malware, and threats separated in time
▪ Host level	▪ All network packets transmitted between hosts are considered	▪ Host and Network (Zhou et al. 2020) – uses statistical and behavioural features, SVM, XGBOOST and Random Forest classifiers, with a confusion classier to detect intrusions	▪ Best performance achieved is 95.2% accuracy (unclear amount of malware samples)

In conclusion, there have been many different approaches (Graph Models, Packet2Vec, Wavelet) using many different aggregating methods (flow, session, connection, host level). The feature map (Sun, Ochiai, and Esaki 2020) and host & network (Zhou et al. 2020) approaches measure performance in accuracy, which, given the fact that most datasets have small amounts of malware samples, is not meaningful. Performance is better measured using precision or an F1 score. The best performing approach is DL-IDS (Sun, Liu, Li, Liu, Lu, Hao, and Chen, 2020), which achieved an overall F1 score of 93.32%. In achieving a dynamic self-learning RL cyber defence with an automated response, higher levels of F1 scores are required. From Sun et al. (2020), there does, however, seem promise in using session aggregation and including an image of the transmitted data within the fingerprint. In addition, there is no implementation of RL learners directly from PCAPs evidenced in Table 1.

Therefore, for this research, the fingerprint will be developed using the session aggregation approach and include a visualisation of the transmitted data, building on the successes achieved by DL-IDS.

2.2 Research that focuses on visual data mining and encoding data into visual information

Visual data mining techniques have become the key adaptation that is applied to data to develop representations for inputs into AI or machine learning algorithms (Wu, Wang, Shu, Moritz, Cui, Zhang, Zhang, and Qu 2021). Visual data mining is already widely used in transformation, assessment, comparison, querying, recommendation, mining and reasoning applications (Wu et al. 2021).

Regarding large datasets, one visual data mining technique, namely pixel-orientated visualisation, has developed a formal representation for encoding data using space-filling techniques (Keim 1996). The primary principle of a space-filling curve is to provide a continuous curve that passes through every point of a spatial region.

One problem, according to Keim (1996), is finding a meaningful arrangement for mapping the pixels because there are many possibilities for arranging the data. Techniques that provide clustering of closely related data are more meaningful and allow for discovering patterns that would otherwise be difficult (Keim 1996). This is vital for the research in this paper, which is aimed at maximising the discovery potential of patterns within network traffic.

Keim (1996) have identified four possible arrangements, line-by-line, column-by-column, Hilbert, and Morton, which are the most appropriate for data with a natural ordering. While line-by-line and column-by-column would provide a space-filling curve, it would result in points being close together in 1D, not being close together after being mapped to 2D. This is not ideal for this research. Hilbert and Morton, on the other hand, map 1D items so that they remain close in 2D. In addition, network packets contained in the network trace PCAP files are generally stored in the order in which they were received. Therefore, network trace data has a natural ordering, which is meaningful for visual data mining.

According to Keim (1996), both the Hilbert and Morton curves have excellent clustering capabilities, because of the resulting 2D space remaining relatively consistent with the 1D input, as depicted in Figure 2. The clustering capabilities provide the most meaningful representations for discovering hidden patterns.

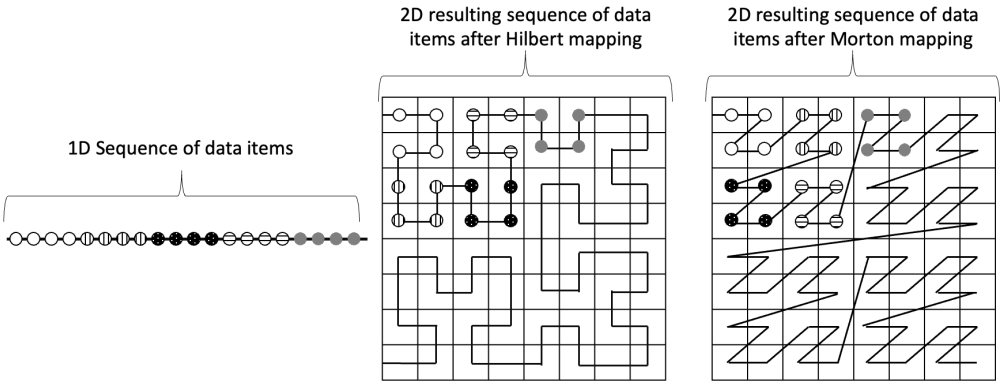


Figure 2: Example 1D mapping to 2D for Hilbert and Morton curves

The Hilbert curve, however, is consistent in that all data items in the 2-dimensional encoding remain close to the data items that were close in the 1-dimensional form, unlike with the Morten curve, as depicted in Figure 2. This property of the Hilbert curve is vital for encoding network session packets.

Therefore, for this research, to ensure that the resulting fingerprint is effective in extracting meaningful information from network packets, the fingerprint will be encoded using the Hilbert curve approach.

3. Fingerprinting Network Sessions

To develop a fingerprint that can be created in real-time as network packets are received, it should be limited to the information to which a firewall has access, since inspecting packets too deeply will add significant delays and will duplicate the functionality of the web application firewall (WAF). At a high level, a firewall has four functions and access to source and destination IP addresses, ports, protocol, associated flags and the transmitted data (Ingham and Forrest 2002), as depicted in Figure 3.

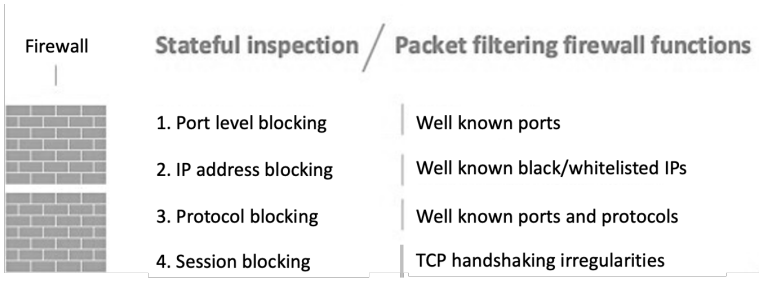


Figure 3: Firewall functions and data access

The methodology to develop a fingerprint based on the information available to a firewall and subsequent cyber threat detection can be structured in the following different steps (steps 4 to 5 are not in the scope of this paper):

1. Collect PCAP data
2. Extract only relevant data from PCAP
3. Formulate a representation to fingerprint unique sessions

4. Train RL to discover and classify cyber threats, with as little information as possible (by rewarding discovery with incomplete fingerprints more) and optimise to maximise performance
5. Detect cyber threats from PCAP test data

3.1 Collect PCAP data

There are several publicly available PCAP data sets; however, a dataset provided by DARPA was selected as the basis for this research. UNSW-NB15 (Moustafa and Slay 2015) made the DARPA data set available in 2015. This data set includes attacks that include, amongst other threats, DoS, worms, backdoors, and fuzzers. These attacks include "zero day attacks". The DARPA dataset contains 100 GB raw network packets (PCAPs). Due to the size of the dataset, and that it contains raw network packets with labels, the DARPA UNSW-NB15 dataset was selected as the optimal dataset for this research. It contains 82 Mil network packets in the training dataset alone. In addition, this dataset contains threat labels for categories of threats rather than actual attacks, which will be more meaningful in training a high-performing threat detection algorithm.

3.2 Extract relevant data from PCAP

The UNSW-NB15 PCAP dataset contains a wide assortment of packets, including IP and address resolution protocol (ARP), higher layer protocols and a wide array of lower layer protocols. Lower layer protocols include TCP, UDP, SCTP, GRE, ICMP, ESP, AH and VRRP. Not all protocols contain source and destination ports, such as all other lower-layer protocols, excluding TCP and UDP. Therefore, as minimum source and destination IP addresses (psrc and pdst in ARP) are collected and where ports are not available, a unique ID to facilitate fingerprinting is selected.

From the research conducted in extracting features from PCAPs, it is clear that there is a lot of meaningful data hidden within the sequence of events. Therefore, in addition to the unique identification for each fingerprint, the protocol discourse and transmitted data are also collected. The protocol discourse is the collection of packets that are sent by each party, represented by the length of each packet, and for TCP packets, including their associated flags. For all other packet types, a default flag is set. The transmitted data is the collective payload in bytes of each packet within a unique session.

Therefore, IP source, IP destination, IP length, TCP flags, source port, destination port, protocol, ARP p-source, ARP p-destination and the raw data are meaningful features for the purpose of this research.

3.3 Formulate a representation to fingerprint unique sessions

Key limitations to consider for formulation of the fingerprint are the following:

1. A Hilbert curve has an output that is always in squares of 2x2, 4x4, 8x8, 16x16, 32x32, 64x64, etc.
2. Neural network models input and output shape need to remain constant; any change will result in retraining since there will be a mismatch in weights.

With the limitations in consideration, the fingerprint design has three distinct sections that have different encoding approaches. These are a unique header, the protocol discourse and the transmitted data.

Each fingerprint section is discussed in more detail below.

3.3.1 Header section

The header of the fingerprint needs to be unique for every session. This is achieved by using the source and destination IP addresses, ports and protocols as depicted in Figure 4.

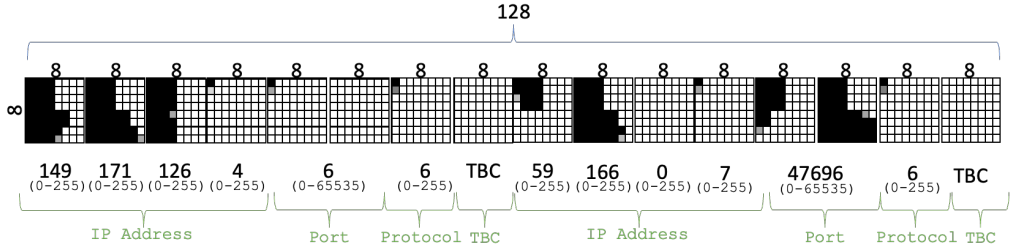


Figure 4: Fingerprint header design

IP version 4 addresses have four sections that range from 0 to 255, each separated by a ".". Using four colours and an 8x8 Hilbert curve, each IP section can be encoded as depicted in Figure 3.

Both TCP and UDP port numbers range from 0 to 65535, which can be encoded using four colours and two 8x8 Hilbert curves. This is achieved by counting 255 in the first Hilbert for every 1 counted in the second Hilbert curve.

Protocols range from 0 to 255 and are encoded in a similar approach as IP sections. The last 8x8 Hilbert curve is reserved for future use and is required to complete the 128 columns required for a 128x128 Hilbert curve, which is used for the transmitted data section. One possible future use considered is to record the frame sizes for IP, TCP, UDP and lower-level protocols in the last 8x8 Hilbert.

3.3.2 Protocol discourse section

The protocol discourse section of the fingerprint needs to portray the exchange of packets between the source and destination hosts for every network session. This is achieved by using the packet length, associated flags and where the packet originates from by using the positive and negative y-axis, as depicted in Figure 5. This visualisation is an adaptation of the typical tornado diagram (Eschenbach 2006).

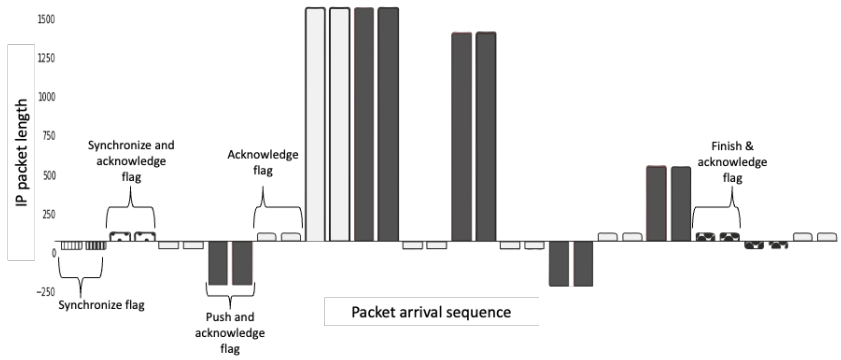


Figure 5: Protocol discourse design

In Figure 5, the protocol exchange is visualised for a TCP session between 59.166.0.7 on port 53421 and 149.171.126.4 on port 80. The exchange is initiated with a request to synchronise, which is acknowledged, after which an initial setup is acknowledged and push and acknowledged before large packets are sent and acknowledged. This is finalised in the end with a finish and acknowledge before closing the exchange.

The fingerprint has enough capacity to capture 128 interactions between two hosts, which can contain multiple flows within the same unique session.

3.3.3 Transmitted data section

The transmitted data section of the fingerprint needs to encode the packet data for all the packets within a unique session, or until the 128x128 Hilbert curve is completed. A fully used Hilbert curve is depicted in Figure 6.

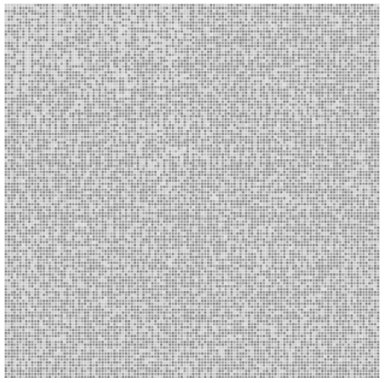


Figure 6: Transmitted data design

Data is transmitted in bytes, which are made up of 8 bits or 0 and 1s. As a result, each byte is converted into decimal range from 0 to 255, which can be encoded into grey scale colours. Therefore, each element of the 128x128 Hilbert curve can depict a byte using grey scale colours.

3.3.4 Proposed dynamic and automated self-learning cyber defence system

Data in the UNSW is extracted in batches for each PCAP separately. Data preparation involves preparing the data for fingerprinting as depicted in Unified Modelling Language diagram below in Figure 7.

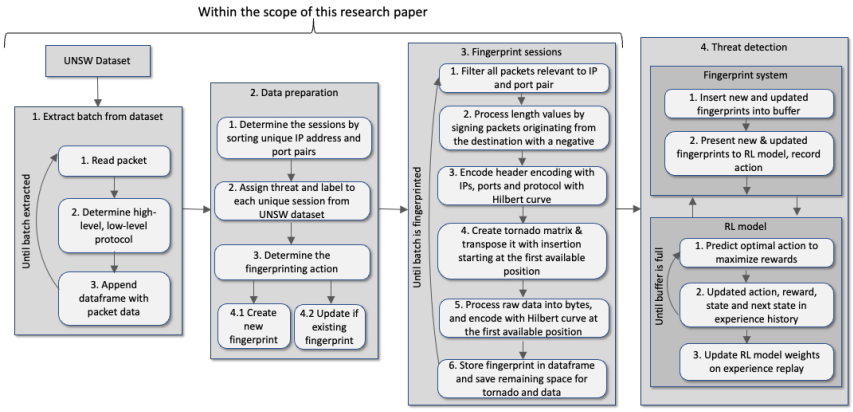


Figure 7: Proposed dynamic and automated self-learning cyber defence system

The fingerprinting system can present fingerprints to the RL model when a unique header is created from the first packet that is exchanged, and then to add new protocol discourse and transmitted data, packet for packet in real-time, or in batches when processing PCAPs.

The RL model predicts the optimal action, which could be assigning a threat or benign classification, or requiring more of the fingerprint to be updated.

The authors of the research at hand plan to train AI and ML models using the fingerprint data as inputs to significantly increase the detection rate of the nine threats within the UNSW dataset, including discovery of "zero-day attacks". The full scope of the proposed self-learning cyber defence system is captured in Figure 7.

4. Results

Using the algorithm and fingerprint system on the UNSW dataset, each unique session can be fingerprinted as benign or malicious. Figure 8 represents a unique fingerprint for a benign network session which contain no malware and Figure 9 represents a unique fingerprint containing malicious shellcode.

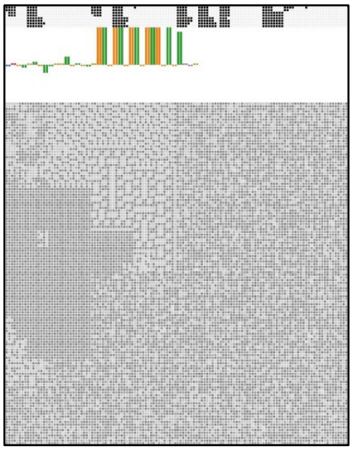


Figure 8: Benign network session fingerprint

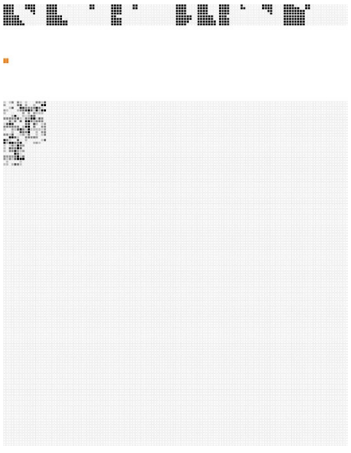


Figure 9: Malicious network session fingerprint

5. Conclusion and future work

The main contribution of this paper is the design of a unique fingerprint by extracting meaningful information from network packets and the fingerprinting system, which is achieved by combining advances in cybersecurity research and in visual data mining. The results in this paper demonstrate that visual ability to discriminate multiple malicious attack types from benign and from one another. Therefore, the results in this paper will lead to more research in RL in the field of cyber security, which will inspire the development of a self-learning dynamic RL cyber defence. Achieving meaningful extraction of information and enabling training of self-learning dynamic RL cyber defence systems, the discovery of undetectable malware, "zero-day attacks" and ransomware, will be possible since fingerprints will significantly simplify the decision boundary for malware detection.

The protocol discourse's unique properties represent the possibility of further study of the possible classification of the application from which the network session was generated. This is significant since there is no approach in place that can accurately classify traffic per application on the open internet accurately for all application in operation today.

Lastly, future work can also be conducted on how to further improve the effectiveness of the fingerprint to contain more information, while improving AI, ML or the effectiveness and performance of RL models.

Author's contributions – The author contributed the entire fingerprint design and subsequent RL model design, including the selection of features for the header, protocol discourse and transmitted data sessions, the encodings as well as the extracting of the relevant features from PCAP files.

Acknowledgements

I would like to thank the following people for their contribution to this research:

- Professor Eloff, for his guidance, expertise, patience, and time.
- My wife and children, for their love and support, that helped me to make this dream a reality.

References

Caminero, G., Lopez-Martin, M. and Carro, B., 2019. Adversarial environment reinforcement learning algorithm for intrusion detection. *Computer Networks, 159*, pp. 96-109.

Du, J., Raza, S.H., Ahmad, M., Alam, I., Dar, S.H. and Habib, M.A., 2022. Digital forensics as advanced ransomware pre-attack detection algorithm for endpoint data protection. *Security and Communication Networks*, *2022*.

Du, Z., Ma, L., Li, H., Li, Q., Sun, G., & Liu, Z., 2018. Network traffic anomaly detection based on wavelet analysis. In *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)* (pp. 94-101). IEEE.

Eschenbach, Ted., 2006. Technical note: constructing tornado diagrams with spreadsheets. *The Engineering Economist, 51*(2), 195-204.

Goodman, E. L., Zimmerman, C., & Hudson, C., 2020. Packet2vec: utilizing word2vec for feature extraction in packet data. *arXiv preprint arXiv:2004.*14477.

Indusface, 15 Malware statistics to take seriously in 2022. Indusface,. https://www.indusface.com/blog/15-malware-statistics-to-take-seriously-in-2022/ Accessed 25 October.

Ingham, K., & Forrest, S., 2002. A history and survey of network firewalls. *University of New Mexico, Tech. Rep*.

Kaloudi, N., & Li, J., 2020. The ai-based cyber threat landscape: a survey. *ACM Computing Surveys (CSUR)*, *53*(1), 1-34.

Kaspersky, What is a Zero-day atack? - Definition and explanation. Kaspersky. https://www.kaspersky.co.za/ resource-center/definitions/zero-day-exploit, Accessed 29 June.

Keim, D.A., 1996. Pixel-oriented database visualizations. *ACM Sigmod Record, 25*(4), pp. 35-39.

Lopez-Martin, M., Carro, B. and Sanchez-Esguevillas, A., 2020. Application of deep reinforcement learning to intrusion detection for supervised problems. *Expert Systems with Applications, 141*, p. 112963.

Malialis, K., 2014. Distributed reinforcement learning for network intrusion response *(Doctoral dissertation, University of York.)*

Moustafa, N., & Slay, J., 2015. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (Milcis)* (pp. 1-6). IEEE.

Nari, S., & Ghorbani, A. A., 2013. Automated malware classification based on network behavior. In *2013 International Conference on Computing, Networking and Communications (ICNC)* (pp. 642-647). IEEE.

Sobers, R., 2022. 89 Must-know data breach statistics [2022]. Varonis, May 2022. https://www.varonis.com/ blog/cybersecurity-statistics, Accessed 29 June. 2022.

Nguyen, T. T., & Reddi, V. J., 2019. Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*.

Sun, P., Liu, P., Li, Q., Liu, C., Lu, X., Hao, R., & Chen, J., 2020. DL-IDS: extracting features using CNN-LSTM hybrid network for intrusion detection system. *Security and Communication Networks*, *2020.*

Sun, Y., Ochiai, H., & Esaki, H., 2020. Intrusion detection with segmented federated learning for large-scale multiple lans. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Velea, R., Ciobanu, C., Gurzau, F., & Patriciu, V. V., 2017. Feature extraction and visualization for network PCAPNG traces. In *2017 21st International Conference on Control Systems and Computer Science (CSCS)* (pp. 311-316). IEEE.

Wu, A., Wang, Y., Shu, X., Moritz, D., Cui, W., Zhang, H., Zhang, D. and Qu, H., 2021. Ai4vis: survey on artificial intelligence approaches for data visualisation. *IEEE Transactions on Visualization and Computer Graphics.*

Zhou, J., Niu, W., Zhang, X., Peng, Y., Wu, H., & Hu, T., 2020. Android Malware Classification approach based on host-level encrypted traffic shaping. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 246-249). IEEE.

S-400s, Disinformation, and Anti-American Sentiment in Turkey

Alex Korb[1], Saltuk Karahan[1], Gowri Prathap[2], Ekrem Kaya[2], Luke Palmieri[2] and Hamdi Kavak[2]

[1] Old Dominion University, Norfolk, VA, USA

[2] George Mason University, Fairfax, VA, USA

rkorb001@odu.edu
skarahan@odu.edu
gprathap@gmu.edu
ekaya4@gmu.edu
lpalmie@gmu.edu
hkavak@gmu.edu

Abstract: As social and political discourse in most countries becomes more polarized, anti-Americanism has risen not only in the Middle East and Latin America but also among the U.S. allies in Europe. Social media is one platform used to disseminate anti-American views in NATO countries, and its effectiveness can be magnified when mass media, public officials, and popular figures adopt these views. Disinformation, in particular, has gained recognition as a cybersecurity issue from 2016 onward, but disinformation can be manufactured domestically in addition to being part of a foreign influence campaign. In this paper, we analyze Turkish tweets using sentiment analysis techniques and compare the model's results to the manual investigation based on qualitative research. We investigate institutional conditions, social and mass media control, and the state of political discourse in Turkey and focus on narratives pertaining to the purchase of S-400 missiles from Russia by Turkey, as well as the actors spreading these narratives, analyzing for popularity, narrative type, and bot-like behavior. Our findings suggest that although anti-American sentiment has held relatively steady in Turkey since 2003, the tightening of control over mass media networks in Turkey and the adoption of conspiratorial rhetoric by President Erdogan and his allies in the AKP from 2014 onward amplified anti-American sentiment and exacerbated negative sentiment on social media by pitting users against one another. This study and its findings are important because they highlight the importance of social and psychological components of cybersecurity. The ease by which disinformation efforts, influence operations, and other "softer" forms of cyber- and information warfare can be carried out means that they will only grow more common.

Keywords: Social Media Analysis, Disinformation, Natural Language Processing, Anti-Americanism, Astroturfing

1. Introduction

Anti-Americanism, defined as "consistent opposition to U.S. policy" (Lawson & Hudson, 2015), is nothing new in NATO countries, but has typically fluctuated depending on the U.S. President and the nature of their presidency. According to the Pew Research Center's Global Indicators Database, the percentage of respondents with a positive view of the U.S. in Germany, France, and Spain fell to about 30-50% from 2001-2008, rose to roughly 50-70% from 2009-2016, and then fell back to 30-50% from 2017-2020. In the Republic of Turkey, however, anti-American sentiment has held steadily negative since 2003, with the percentage of respondents with positive views of the U.S. never exceeding 30% (Pew Global Indicators Database, 2021). Since 2014, anti-American sentiment has been amplified because of pervasive Turkish control over mass media, and subsequent attempts to control social media. This paper examines the role of disinformation in social media in perpetuating anti-American sentiment, especially in the case of the S-400 purchase. It defines and examines disinformation and discusses the information environment in Turkey. It examines Turkey's tightening control over social media sources combined with the increasingly conspiratorial nature of political discourse since 2014. It analyzes political rhetoric in social media, using the purchase of S-400 missiles and subsequent events surrounding them as a case study, incorporating sentiment analysis data to analyze each event's impact on public opinion of the U.S.

2. Background

2.1 Disinformation and the Information Environment

With cyberspace now being widely recognized as a fifth domain of operations (Welch, 2011), disinformation has been effectively weaponized by state and non-state actors to pursue strategic goals. Often referred to as "fake news," disinformation refers to the production and targeted dissemination of misleading content to "generate profits, pursue political goals, or maliciously deceive" (Humprecht, Esser, and van Aelst, 2020). Social media bots, troll armies, and astroturfing techniques feature very heavily in disinformation campaigns (Kirdemir, 2020).

Disinformation campaigns are not a new phenomenon; during the Cold War, the Soviet Union made use of "active measures" in order to discredit Western politicians and inflame tensions both within target states and between them (Mahairas and Dvilyanski, 2018). However, social media platforms, Twitter among them, have received attention as vectors for disinformation as early as 2010 (Chamberlain). Twitter has received particular attention due to the brevity of posts or "tweets" at 280 characters, the follower/retweet system allowing a tweet or hashtag to rapidly grow in popularity, and the application programming interface (API) allowing for automated posting and datamining (Chamberlain, 2010).

The effectiveness of a disinformation campaign or influence operation depends heavily on the vulnerability of an actor's information ecosystem. The information ecosystem, or information environment, is a broad analytical paradigm that refers to how information is distributed, who has access to this information, who has ownership of infrastructure and channels, how information is used, who or what influences the shape of information, what information is trusted, and the impact of information (Susman-Peña, 2015). Political polarization and hyper-partisanship, a hostile state of political discourse, and regulatory capture of media news networks by the state are all characteristics of a vulnerable information ecosystem, which can then be polluted by disinformation (Kirdemir, 2020).

Although the proliferation of disinformation on social media by state and non-state actors is well-documented, less clear is the relationship between authoritarian or hybrid regimes and social media platforms and their users. Russia continues to make use of active measures against foreign and domestic opponents, utilizing bot farms to spread pro-Kremlin propaganda on Twitter (Abrams, 2016), but Russia requires all social media data on Russian citizens to be stored within its borders and bans members of the military from holding social media accounts (Allen and Moore, 2018), and as a result of the Russian invasion of Ukraine, the majority of social media activity in Russia now takes place on Telegram (Roach, 2022), despite efforts by Russia to ban it. Meanwhile, efforts to study Chinese social media activity are stymied by a strict regime of censorship and monitoring referred to as the "Great Firewall of China," but occasional glimpses into the Chinese government's social media activities reveal a concentrated effort to distract and dilute public discourse by hiring men and women to post messages defending the government and its leaders, referred to as "50c party posts" due ostensibly to being paid 50 cents per post (King, Pan, and Roberts, 2017).

The authoritarian turn taken by Turkish President Recep Tayyip Erdogan allows for easy insight into the practices of anocratic and autocratic regimes on social media. Although Twitter is subject to stringent censorship laws in Turkey, and Turkey even went so far as to temporarily block Twitter in March 2014, Twitter's openness relative to other social media platforms allows for better observation of practices and actors on Turkish social media. Notably, Turkish censorship and practices towards its opposition resemble that of Russia and China, offering insight into the social media practices of the latter two states towards their domestic audiences that otherwise would not be available. Because of this, studying Turkey's social media environment offers insight into the ways that anti-American sentiment is disseminated and promoted by government entities and the ways in which this sentiment affects both the relationship between public officials and domestic audiences and relations with the U.S.

2.2 Turkey's Information Environment

Turkey's information ecosystem has proven vulnerable to disinformation, and domestic and foreign actors have taken advantage. The Turkish news media has undergone regulatory capture by Erdogan and the Justice and Development Party (AKP). Prior to 2013, Turkish media was relatively open, although there were instances of self-censorship over topics considered taboo, such as the family of Erdogan, who was prime minister at the time. The Gezi Park Protests of 2013 heralded a significant change in Turkey's relationship with its news media (Weise, 2018) and opposition. As the protests were anti-authoritarian, pro-Western, and most of all anti-Erdogan, they were squashed by the police force - but mainstream news networks did not cover the protests at all, with CNN Turk instead broadcasting a documentary about penguins, a decision that was received with much ridicule from journalists and news networks elsewhere (Oktem, 2013).

The nature of the Gezi Park protests led Erdogan to adopt increasingly adversarial language concerning NATO and the West from 2014 onward, language that took on a conspiracist aspect when Erdogan mentioned a "üst akıl (mastermind)" a sinister conspiracy aimed at undermining or destroying Turkey (Akyol, 2016). This conspiracist language, as well as the tendency to designate foreign enemies as being behind Turkey's domestic

woes, became more prominent in the aftermath of the coup attempt in July 2016. Using the coup attempt as a pretext, Erdogan both intensified his rhetoric against the United States and NATO and tightened his grip on major media news networks in Turkey, purging journalists and network owners unfriendly to his party (AKP) and replacing them with family members and political allies; in August 2018, the Turkish government was estimated to control 85% of all news media networks in Turkey, according to Turkish-French journalist Erol Onderoglu (Weise, 2018). The majority of mass media in Turkey began loudly parroting Erdogan's accusations; as such, the Turkish government has been actively participating in the production and dissemination of disinformation over mass media, exacerbating the existing political polarization and toxic political discourse in Turkey (Kirdemir, 2020).

Erdogan's influence over media has not been limited to mass media; the AKP has also gradually tightened restrictions on social media since the coup attempt, with the latest attempt being a law to punish those who spread what Turkey defines as disinformation with up to five years in prison (Wilks, 2021). In addition, the AKP has, in the past, employed a troll army of roughly 6,000 users, known as “AK Trolls”, to harass opponents on Twitter (Benedictus, 2016).

3. Methodology

We identify a pivotal event in U.S.-Turkey relations, the purchase of the S-400 missile system from Russia, and analyze the social media narrative around the purchase and later events involving the S-400s. To do this, we collect and translate large volumes of Turkish tweets discussing both the S-400 purchase and the U.S. in general. We then use Afinn sentiment analysis to determine the average sentiment score for all tweets collected on a given day. Following this, we perform qualitative analysis on selected individual tweets for their narrative content and identify the actors creating these tweets based on class (government official, academic, journalist, etc.) and authenticity (human, bot, fake account, etc.).

3.1 The S-400 Case Study

On July 12, 2019, Turkey received its first shipment of Russia's S-400 air defense system, which President Erdogan had signed a deal to purchase in September 2017. Although later events, such as U.S. President Joe Biden's recognition of the Armenian Genocide on April 24, 2021 would exceed it in negative sentiment, this purchase and delivery was the culmination of steadily degrading relations between Turkey and the United States since the attempted coup on July 15, 2016 in Turkey. In the aftermath of the delivery, government officials and pro-Erdogan actors sought to justify the purchase, citing the necessity of the S-400 in vague terms, deflecting attention away from the decision by singling out foreign and domestic enemies and taking advantage of the poor domestic information environment in order to spread disinformation. These justifications, especially the singling out of political foes through disinformation, inflame the already-existing anti-American sentiment in Turkey.

The S-400 purchase is a persistent issue in U.S.-Turkey relations, and as such, we analyze sentiment data and media narratives around three time periods in particular: the month of July, 2019, during which Turkey received the S-400 missiles and was removed from the F-35 project as a result; February 20, 2020, during which time Turkish positions in Syria were attacked by Syrian and Russian jets, and Turkey demanded Patriot missiles from the U.S. despite already having the S-400s; and the month of December, 2020, in which the U.S. sanctioned Turkey for their purchase of the S-400s.

3.2 Dataset

Narrative examination of Turkish Twitter activity surrounding the S-400s involved the manual collection, translation, and analysis of tweets during each time period. These tweets were collected at varying thresholds of popularity (fewer than 100 favorites, more than 1,000 favorites, more than 5,000 favorites, more than 25,000 favorites, etc.) and accounts for tweets made by both government officials, AKP members, academics, journalists, and ordinary members of the public. By examining the language used and identifying keywords, we categorize tweets as promoting "national defense necessity", "national pride", or "attacking opponents" narratives. These categories are not mutually exclusive as some users like former Ankara mayor Gokcek used a combination of these narratives (Gökçek, 2019)

The official position taken by the Turkish government, both before and after the delivery, is that the S-400 missiles are necessary for Turkey's national security. This was the position taken by Turkish Interior Minister

Süleyman Soylu in the pro-government newspaper Yeni Şafak, claiming that Turkey could be attacked by long-range missiles (Ahval News, 2019). Similarly, Turkey's Defense Minister Hulusi Akar described the position as "national matter" and an exercise of "sovereignty rights"(Daily Sabah, 2021). As such, we find that most Turkish public officials push the national defense necessity narrative in their tweets. At the lower levels of the Turkish public, however, attacks on foreign and domestic enemies are more prominent.

4. Results

Thanks to the ML supported anaylsis, we were able to categorize the dominant narratives in anti-American public opinion. We find that national pride and attacks on foreign and domestic opponents were the dominant narratives surrounding the S-400 purchase and related events, and these narratives were predominant among sockpuppet accounts as well. However, we also found significant anti-Erdogan activity on February 20, and the tweets and accounts do not easily fit the qualitative analysis since they do not defend Erdogan at all, although they still contribute to the sentiment score. This can be viewed as an indicator supportive of the observation of domestic political alliances against the U.S. within Turkey.

As seen in the graphs below, we were able to look into three specific events the months where these events took place. The first one is July 2019 when the S-400s have been delivered and Turkey was removed from the F-35 program, the second month is February 2020 when Turkey asked the U.S. to station Patriot missiles on its southern border, and the third period is December 2020 when the United States imposed sanctions on Turkey in response to Turkey's purchase and continued ownership of the S-400s. In each of these periods, the discourse in twitter had changing sentiment scores, and the combination of narratives as well as users (human or bot activity) behind the tweets changed in terms of their dominance.

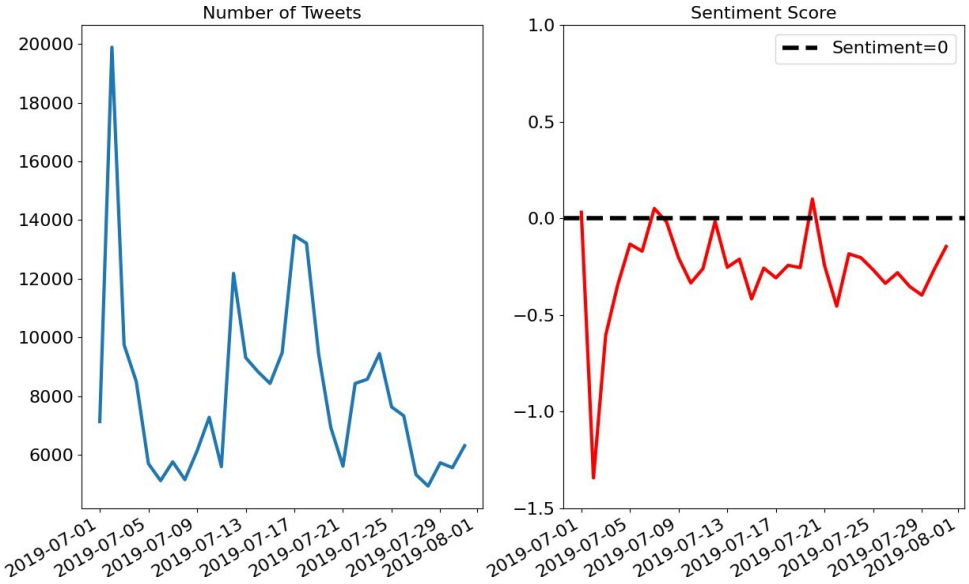


Figure 1: Total number of tweets collected per day and Sentiment Score of tweets dealing with U.S. related topics per day in July 2019

When the S-400s arrived in Turkey on July 12, 2019, national pride as a narrative or justification for their purchase became prominent in addition to the national security narrative and recrimination of perceived enemies. Journalist and AKP deputy Şamil Tayyar extolled the historic nature of the purchase, claiming that Turkey will have had "a military promotion" and urging the public to be vigilant against "traitors" (Tayyar, 2019). There is also evidence for the presence of fake accounts pushing a pro-Erdogan viewpoint. "Hrant Enveryan" identifies himself as an Armenian living in Turkey and an "academic" from Bogazici University. However, Bogazici University has no record of any Hrant Enveryan, and Enveryan has since deleted any mention of Bogazici University from his Facebook and Twitter accounts; furthermore, his profile picture is that of Turkish-Armenian journalist Hrant Dink, who was famously assassinated in January 2007 (Medium, 2020). Enveryan (2019) attacked the CHP for speaking out against the purchase, claiming that the CHP were "paranoid" and against the purchase "from day one", but that "if it was statues instead of #S400, the CHP would have declared it a holiday."

A week after the delivery of the S-400s, the U.S. removed Turkey from the F-35 program for as long as the S-400s remained on Turkish soil. Turkish social media users reacted negatively, but some bot activity was detected as well. The account of Medya Muhtari (``Media Manager''), a user supposedly offering ``digital news and commentary'' (Dijital Haber ve Yorum), was created in August 2016, and has a personal website whose domain has expired, mediamuhtari.com. The account has a score of 2.2 on Botometer; in addition, it has a 54% bot score according to BotSight, indicating a ``bot with human characteristics''. Medya Muhtari commented on the U.S. decision to remove Turkey from the F-35 program supposedly after Turkey had already ordered most or all of them:

> *"The expected sanctions on the S400s have been announced! Trump: Because Turkey bought S400, we will punish by not giving more than 100 F35 planes!" *Turkey's orders through 2024 is a total of 100 F-35s! Is there a better USA than this!" (Muhtarı, 2019)*

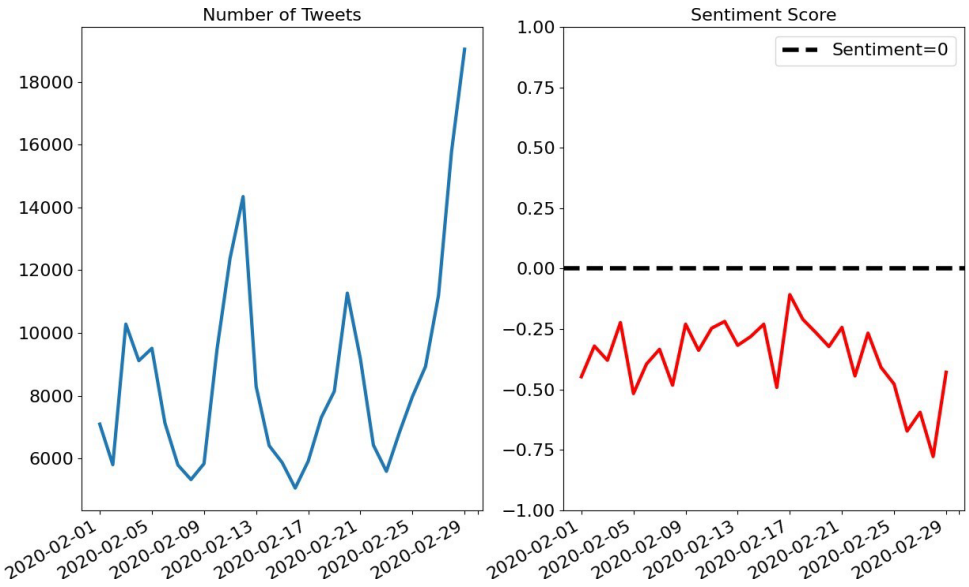


Figure 2: Total number of tweets collected per day and Sentiment Score of tweets dealing with U.S. related topics per day in February 2020

On February 20, 2020, with hostilities escalating between Turkish troops in Syria's Idlib governorate and Syrian government forces, Turkey asked the U.S. to station Patriot missiles on its southern border (Hacaoglu, 2020). This request was met with ridicule from Turkish Twitter users; journalist Levent Gültekin (2020) summarized the criticisms of Erdogan's opponents towards this move and in general:

> *"Pay a billion dollars 6 months ago and buy S400 from Russia and now order Patriot from the USA. This decision is without foresight, it is so irrational, it is to maintain a foreign policy like a chicken with its head cut off… and they shamelessly shout at us that 'we know what we are doing'."*

Pro-Erdogan accounts were not completely absent, however, as one of the very first replies to Gültekin's tweet came from Emine Celik, an account with a rating of 2.3 on Botometer and a 53% human score on BotSight, indicating a "human with some bot-like characteristics". Çelik came out in defense of Turkish defense minister Hulusi Akar, claiming that the news was a hoax (2020). There was bot activity around this topic as well, but unusually, not all bot activity was pro-Erdogan. Ekrem Koca, an account with a 3.1 rating on Botometer and a 75% bot rating on BotSight, indicating a likely bot, came out against Turkey's foreign policy choices:

> *"How can you spend a fortune on s400s and then now buy the Patriot from the U.S. This is not acceptable foreign relations. You can't be enemies with your friends of yesterday. This is absurd. It is a pity, the people of this country will pay for these mistakes." (Koca, 2020)*

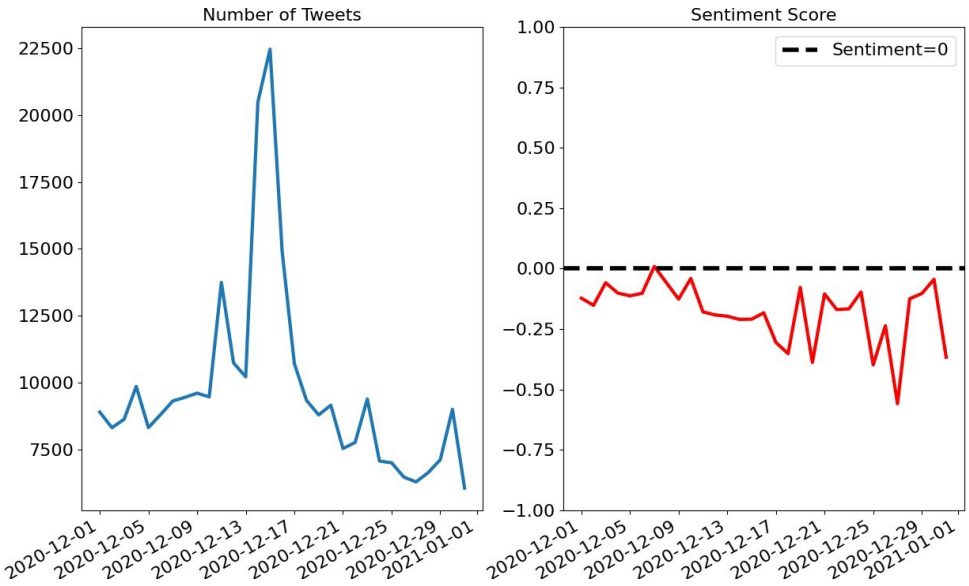


Figure 3: Total number of tweets collected per day and Sentiment Score of tweets dealing with U.S. related topics per day in December 2020

On December 14, 2020, the United States imposed sanctions on Turkey's weapons procurement sector in response to Turkey's purchase and continued ownership of the S-400s. The response from both government officials and the general public was negative; Turkish presidential spokesperson Fahrettin Altun condemned the sanctions in a four-tweet thread, blaming the United States for appeasing shadowy ``anti-Turkey lobbies'' instead of ``fighting terrorists'' and declaring that Turkey "takes all necessary steps to ensure its national security and to promote regional and global security" (2020). Bayram Şenocak, a member of the AKP's Central Decision Board (Merkez Karar Yonetim Kurulu), condemned the United States for its desire to undermine Turkey's defense industry, without the S-400s factoring into it:

> *"S-400 is an excuse. The main reason for the sanctions decision of the USA is that the share of locality in the defense industry in Turkey has increased from 20 percent to 70 percent, and that the effectiveness of our domestic weapons in Libya, Syria and most recently in Karabakh has attracted the commercial interest of the countries." (Şenocak, 2020)*

5. Conclusion and Future Work

Our study used a combination of AI/ML and manual investigations to determine certain dates/times and narrative types in feeding or exacerbating public opinion in social media. Within a specific case, we were able to determine the trend in sentiments and emotions in a U.S. ally. This study argues that further AI/ML techniques and a combination of them with consequent manual analysis and scaling of the study to include other U.S. allies will provide a tool for understanding to what extent each narrative shapes anti-American sentiment in U.S. allies. Furthermore, the narratives are generally indicator of domestic groups or political actors causing anti-American sentiment and such an AI/ML supported analysis can be utilized to understand the specific agendas created by each domestic political actor in a country.

Turkish sentiment on Twitter has been consistently negative, with notable dips whenever an incident with the United States occurs and users react to it. The conspiracist angle, which remained on the fringe until Erdogan adopted it from 2014 onward, has filtered down to other members of the AKP and the general public, with Turkish Twitter users regularly accusing the domestic opponents, the United States, and NATO of trying to undermine Turkey. The national defense necessity narrative is much-promoted by the Turkish government and was initially parroted by Erdogan's supporters on Twitter, but has become less common outside of the government circles since the February 20 request for Patriot missiles. The national pride narrative is still present, but has not been associated with the S-400 missiles themselves since the sanctions imposed on December 14.

Bot activity is rampant, as are fake accounts created to promote pro-Erdogan messages. At the same time, however, the backlash to the request for Patriot missiles demonstrates that opposition to Erdogan has not been fully snuffed out on social media. Rather, the arguments between pro- and anti-Erdogan users and bots likely

contribute to the consistent negative sentiment scores for Turkish tweets per day. Both the arguments and the negative sentiment scores are a product of the poor information environment in Turkey; although Erdogan, try as he might, cannot exert as much control over social media as he can over mass media news networks, the consequences of his attempts to do so, as well as the AKP's control over mass media and the extreme state of political polarization and political discourse in Turkey, can be seen in how Turkish Twitter users interact with one another.

This paper has provided a comparatively small glimpse into the larger issue of disinformation and anti-American sentiment in Turkey. Avenues for future research include expanding the scope of study beyond the dates provided and even beyond the S-400s case and studying the presence and effects of foreign disinformation narratives and influence operations used in Turkey. This study can also be used as a model for similar studies in other U.S. allies where anti-Americanism is fed by several domestic actors and has been expressed in different narratives. Understanding of the actors and the narratives is crucial in crafing relevant foreign policy tools and the response to changes in overall public opinion shaped by misinformation and disinformation campaigns.

Acknowledgements

This work is funded by the Commonwealth Cyber Initiative program: The Role of Cybersecurity in the Spread of Misinformation and Disinformation. Award# H-4Q21-015.

References

Abrams, S. (2016) “Beyond Propaganda: Soviet Active Measures in Putin's Russia,” Connections, 15(1). Available at: https://www.jstor.org/stable/26326426.

Akyol, M. (2016) *The tin-foil hats are out in Turkey*, *Foreign Policy*. Available at: https://foreignpolicy.com/2016/09/12/the-tin-foil-hats-are-out-in-turkey/ (Accessed: October 5, 2022).

Allen, T.S. and Moore, A.J. (2018) “Victory without casualties: Russia's information operations,” The US Army War College Quarterly: Parameters, 48(1). Available at: https://doi.org/10.55540/0031-1723.2851.

Altun, F. (2020) *The decision by the United States to sanction Turkey over its purchase of the S-400...* Twitter. Available at: https://twitter.com/fahrettinaltun/status/1338588508483690496 (Accessed: October 5, 2022).

Benedictus, L. (2016) *Invasion of the troll armies: from Russian Trump supporters to Turkish state stooges*, *The Guardian*. Guardian News and Media. Available at: https://www.theguardian.com/media/2016/nov/06/troll-armies-social-media-trump-russian (Accessed: October 5, 2022).

Celik, E. (2020) *Pardon Siz Savunma Bakanlığından Daha mı Iyi Bileceksiniz...* Twitter. Available at: https://twitter.com/senkimsinyaRTE/status/1230544258840629250 (Accessed: October 5, 2022).

Chamberlain, P.R. (2010) “Twitter as a Vector for Disinformation,” Journal of Information Warfare, 9(1). Available at: https://www.jstor.org/stable/26480487.

Daily Sabah (2021) Turkey exercised sovereignty rights with S400s, Akar says, The Daily Sabah, Available at: https://www.dailysabah.com/politics/diplomacy/turkey-exercised-sovereignty-rights-with-s-400-akar-says

Enveryan, H. (2019) *CHP Genel Başkan Yardımcısı ünal çeviköz...* Twitter. Available at: https://twitter.com/enveryan/status/1131546160630571008 (Accessed: October 5, 2022).

Global indicators database (2021), Pew Research Center's Global Attitudes Project. Available at: https://pewrsr.ch/3OZQLN1 (Accessed: October 5, 2022).

Gökçek, İ.M. (2019) *Emekli tümgeneral Beyazit karakaş'dan S-400...* Twitter. Available at: https://twitter.com/06melihgokcek/status/1125174808994623488 (Accessed: October 5, 2022).

Gültekin, L. (2020) *6 ay önce Milyar Dolar ödeyip rusya'dan S400 al şimdi de Abd'den Patriot Siparişi et...* Twitter. Available at: https://twitter.com/acikcenk/status/1230517625039405062 (Accessed: October 5, 2022).

Hacaoglu, S. (2020) *Turkey seeks U.S. patriot missiles to deter Russia in Syria*, *Bloomberg*. Available at: https://www.bloomberg.com/news/articles/2020-02-20/turkey-asks-u-s-for-patriot-missiles-to-deter-russia-in-idlib (Accessed: October 5, 2022).

Humprecht, E., Esser, F. and Van Aelst, P. (2020) “Resilience to online disinformation: A framework for cross-national comparative research,” *The International Journal of Press/Politics*, 25(3), pp. 493–516. Available at: https://doi.org/10.1177/1940161219900126.

King, G., Pan, J. and Roberts, M.E. (2017) “How the Chinese government fabricates social media posts for strategic distraction, not engaged argument,” American Political Science Review, 111(3), pp. 484–501. Available at: https://doi.org/10.1017/s0003055417000144.

Kirdemir, B. (2020). *Exploring Turkey’s Disinformation Ecosystem: An Overview*. Centre for Economics and Foreign Policy Studies. Available at: http://www.jstor.org/stable/resrep26087

Koca, E. (2020) *Sen Kalk Dünyanın Parasına rusya'dan S-400 siparişi ver, şimdi de Abd'den...* Twitter. Available at: https://twitter.com/EkremKocaEsk/status/1230601490294870016 (Accessed: October 5, 2022).

Lawson, C.W. and Hudson, J. (2015) "Who is Anti-American in the European Union?," *SAGE Open*, 5(2), pp. 1–15. Available at: https://doi.org/10.1177/2158244015584163.

Mahairas, A. and Dvilyanski, M. (2018) "Disinformation – Дезинформация (Dezinformatsiya)," The Cyber Defense Review, 3(3). Available at: https://www.jstor.org/stable/26554993

Muhtarı, M. (2019) *Ve S400 konusunda, abd'den beklenen yaptırım geldi!...* Twitter. Available at: https://twitter.com/MediaMuhtari/status/1151201325436329986 (Accessed: October 5, 2022).

Nielsen, F. (2016) *Fnielsen/afinn: AFINN sentiment analysis in Python*, *GitHub*. Available at: https://github.com/fnielsen/afinn (Accessed: October 5, 2022).

Oktem, K. (2013) *Why Turkey's mainstream media chose to show Penguins rather than protests*, *The Guardian*. Guardian News and Media. Available at: https://www.theguardian.com/commentisfree/2013/jun/09/turkey-mainstream-media-penguins-protests (Accessed: October 5, 2022).

Roach, S. (2022) In Russia, telegram has become the messaging app of choice. Protocol. Available at: https://www.protocol.com/bulletins/telegram-whatsapp-meta-russia (Accessed: November 17, 2022).

Şenocak, B. (2020) *S-400 bahanedir. Abd'Nin Yaptırım Kararının Asıl Gerekçesi...* Twitter. Available at: https://twitter.com/bayramsenocak/status/1338595873048186885 (Accessed: October 5, 2022).

Susman-Peña, T. (2015) *Why Information Matters: a Foundation for Resilience*, *UNHCR Innovation*. Available at: https://unhcr.org/innovation/wp-content/uploads/2017/10/150513-Internews_WhyInformationMatters.pdf (Accessed: October 6, 2022).

Suspicious twitter accounts claiming to be Armenians from Turkey spur nationalist rhetoric (2020) Medium. DFRLab. Available at: https://medium.com/dfrlab/suspicious-twitter-accounts-claiming-to-be-armenians-from-turkey-spur-nationalist-rhetoric-2ff7eb883576 (Accessed: October 5, 2022).

Tayyar, Ş. (2019) *S-400 Sevkıyatı Başladı. Tarih Yeniden Yazılıyor...* Twitter. Available at: https://twitter.com/samiltayyar27/status/1149652257853460484 (Accessed: October 5, 2022).

Turkey could deploy S-400 missiles in Libya, columnist says (2020) *Ahval News*. Available at: https://ahvalnews.com/turkey-s-400/turkey-could-deploy-s-400-missiles-libya-columnist-says (Accessed: October 5, 2022).

Weise, Z. (2018) *How did things get so bad for Turkey's journalists?*, *The Atlantic*. Atlantic Media Company. Available at: https://www.theatlantic.com/international/archive/2018/08/destroying-free-press-erdogan-turkey/568402/ (Accessed: October 5, 2022).

Welch, L.D. (2011) *Cyberspace: the fifth operational domain*, *Institute for Defence Analyses*. Available at: https://www.ida.org/-/media/feature/publications/2/20/2011-cyberspace---the-fifth-operational-domain/2011-cyberspace---the-fifth-operational-domain.ashx (Accessed: October 6, 2022).

Wilks, A. (2021) *Turkey's plans for new social media restrictions threaten five years in prison for spreading fake news*, *Al-Monitor*. Available at: https://www.al-monitor.com/originals/2021/08/turkeys-plans-new-social-media-restrictions-threaten-five-years-prison-spreading (Accessed: October 5, 2022).

Detecting and Tracking Hypersonic Glide Vehicles: A Cybersecurity-Engineering Analysis of Academic Literature

Yee Wei Law[1], John Joshua Gliponeo[1], Dilpreet Singh[1], John McGuire[1], Jiajun Liang[1], Sook-Ying Ho[1] and Jill Slay[1,2]

[1]UniSA STEM, University of South Australia, Australia
[2]SmartSat Cooperative Research Centre, Australia
yeewei.law@unisa.edu.au

Abstract: Hypersonic vehicles are vehicles travelling faster than Mach 5 (five times the speed of sound). Hypersonic technologies have existed since the end of the 1950s, but recent developments of defence applications have led to their resurgence. Hypersonic weapons can be hypersonic (powered) cruise missiles or hypersonic glide vehicles (HGVs). The near-space trajectories of HGV, combined with their superior manoeuvrability, enable HGVs to evade existing space and terrestrial sensors used to track ballistic missiles, posing an immediate threat to today's radar networks and making HGVs well-suited for intercontinental (> 5500 km) targets. Securing HGV detection and tracking systems is of great interest to at-risk nations and cybersecurity researchers alike. However, like hypersonic flight technologies, HGV defence technologies are heavily guarded secrets. The shortage of public-domain information did not stop academia from proposing various detection and tracking schemes, but a reasonable question is: "How credible and useful is current public-domain information, including academic publications, on HGV detection and tracking for academic researchers to base their cybersecurity research on?" To answer this question, we scanned and critically reviewed public-domain literature on HGV detection and tracking. We then identified ambiguities and knowledge gaps in the literature. In this paper, we provide a concise version of our multivocal literature review and an analysis of the identified ambiguities and knowledge gaps in our attempt to answer our earlier question.

Keywords: Hypersonic glide vehicle, Hypersonic and Ballistic Tracking Space Sensor, National Defence Space Architecture, satellite constellation, small target detection, target tracking

1. Introduction

Hypersonic weapons can be classified into two main types (Van Wie, 2021): (i) hypersonic (powered) cruise missiles (HCMs), and (ii) hypersonic boost-glide systems (HGVs). HCMs are multistage vehicles that use solid rocket propulsion to accelerate from launch to a speed enabling a scramjet-powered sustainer stage to take over powering the remainder of the flight. HGVs are also multistage vehicles typically powered by solid rockets, but they are simpler by construction. The launch stage accelerates the glide stage to hypersonic speeds at the edge of space, and the glide stage then glides unpowered to its target. While HCMs are for medium-range (1000-3000 km) targets, HGVs are for intercontinental (> 5500 km) targets.

Detecting the launch of an HGV is already within some governments' capability, but the near-space trajectories of HGVs combined with their superior manoeuvrability enable HGVs to evade existing space and terrestrial sensors designed to track ballistic missiles (Sherman, 2022); see **Figure 1**. HGVs however emit strong thermal radiation in the infrared (IR) spectrum when hypersonic gliding raises their surface temperature to several thousand Kelvins (Tracy and Wright, 2020). Thus space-borne and persistent electro-optical/infrared (EO/IR) sensing is widely recognised as the alternative to ground-based radars for HGV detection.

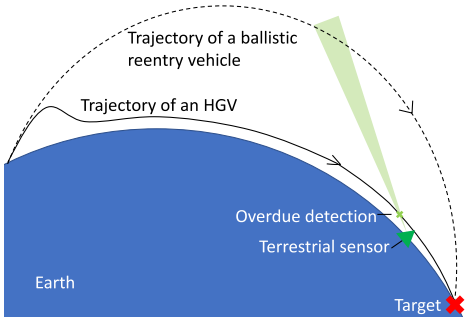


Figure 1: HGV trajectories are not conducive to early detection by terrestrial sensors.

Perhaps the best-known example of a space-borne persistent infrared sensor is the Hypersonic and Ballistic Tracking Space Sensor (HBTSS) envisioned by the U.S.' Missile Defense Agency (MDA) and Space Development Agency (SDA). HBTSS satellites sit in low-Earth orbits (LEOs) with a medium field of view and form a part of the Next-Generation Overhead Persistent Infrared (Next-Gen OPIR) multilayered satellite constellation (Strout,

2019). The Next-Gen OPIR constellation is designed to augment and eventually replace the U.S. Air Force's Space-Based Infrared System constellation, a legacy satellite-based early warning missile defence system (Munoz, 2021). As a vehicle passes from one HBTSS' "field of regard" to another's, tracking information is transmitted by the former to the latter in real time. Integration with terrestrial sensors and high-speed data sharing – potentially through optical links – enable continuous tracking of long-range missiles from launch to impact; this is known as *cradle-to-grave* or *birth-to-death* tracking. The U.S.' blueprint for enabling this level of tracking is called the National Defence Space Architecture (NDSA).

Securing an HGV detection and tracking system, within or without the NDSA context, poses not only significant cyber-engineering challenges, but also important scientific problems. The sensitive nature of these problems means researchers investigating these problems must cope with a dearth of critical information on HGV-related technologies. There is nevertheless a steady production of academic literature on HGV detection and tracking. To assess the credibility and usefulness of this literature for analysing the attack surface and security posture of an HGV detection and tracking system, we ask this question:

> "How credible and useful is current public-domain information, including academic publications, on HGV detection and tracking for academic researchers to base their cybersecurity research on – what are the ambiguities and knowledge gaps?"

This paper represents our first attempt at answering the question above, and we are potentially the first to critically explore this topic. Hence, **our contribution** through this paper is unprecedented multidisciplinary insight on the viability of cybersecurity research in the context of HGV detection and tracking relying only on public-domain information.

In what follows, we will discuss in Sec. 2 the aspects of the NDSA that are relevant to HGV detection and tracking; in Secs. 3-4 the two main sensing modes; in Sec. 5 aerodynamic modelling as a prerequisite for tracking; and in Sec. 6 tracking algorithms. Each of Secs. 3-6 contains a subsection discussing ambiguities and knowledge gaps in the literature. Sec. 8 summarises our findings and provides a concluding note on our future directions.

2. National Defense Space Architecture (NDSA)

As a blueprint for a proliferated mega-constellation from LEOs to geosynchronous orbits (GEOs), the NDSA consists of seven logical layers (SDA, 2021). Going bottom-up,

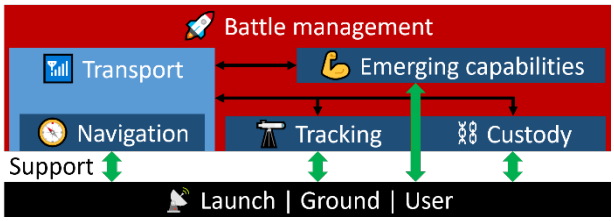


Figure 2: The seven logical layers of NDSA.

The Support Layer provides a common, resilient ground support infrastructure necessary to underpin the space-based capabilities of the other layers to disseminate, process and leverage data.

The Transport Layer provides a low-latency communication backbone, in the form of a mesh network of hundreds of satellites, connected through optical links. Network connectivity extends to existing tactical data links.

The Navigation Layer provides positioning, navigation and timing (PNT) services. By sitting inside the Transport Layer (see **Figure 2**), the Navigation Layer enables the satellite mesh network to beam PNT signals down as navigation messages to warfighters over existing tactical data links.

The Tracking Layer supports space-based sensing and data fusion across sensors and orbital regimes, and thereby provides from LEOs global indications, tracking, and targeting of missile threats, including HGVs.

The Custody Layer maintains "target custody" and provides the ability to detect and track time-sensitive mobile targets on the Earth's surface (ground and sea) from LEOs. It is focused on deriving weapons-quality geolocations and tracks.

The Emerging Capabilities Layer supports new mission concepts in areas such as space situational awareness and environmental monitoring.

The Battle Management Layer provides automated space-based battle management through command control, mission processing, tasking, and information dissemination to support time-sensitive kill chain closures.

While a holistic approach to securing HGV detection and tracking is analysing all seven layers of the NDSA and their interactions, the focus here is the Tracking Layer, the outputs of which are used by *glide phase interceptors* to close the sensor-to-shooter loop or kill chain. To support beyond-line-of-sight targeting of HGVs by the interceptors, the Tracking Layer needs to generate tracking data of "fire control quality" (Sherman, 2022).

For HGV detection, two sensing modes are of concern: EO/IR sensing and radio detection and ranging (radar), which are discussed in Sec. 3 and Sec. 4 respectively.

For HGV tracking, stochastic estimation/filtering techniques are relevant. These techniques however require a kinematic model of the target. This model typically takes the form of a discrete-time state equation, which allows the target's coordinates and velocity at the next time interval to be predicted based on the measurement/observation at the current time interval. The superior manoeuvrability of HGVs rules out the use of state equations based on simplified kinematic models such as the constant velocity model, the constant acceleration model, the constant turning model and the Singer model (Singer, 1970). The kinematic models used for target tracking do not need to be as detailed as those used for guidance and control, but they do need to take *hypersonic aerodynamics* into account. Following this logic, Sec. 5 discusses aerodynamics modelling before tracking algorithms are discussed in Sec. 6.

3. EO/IR sensing

Detecting HGVs by EO/IR sensing from GEOs is challenging because HGVs appear to be one order of magnitude dimmer than the usual targets (Strout, 2019). Moving sensors from GEOs to LEOs helps, but the associated *signal-to-clutter ratios* (see Definition 1) are still lower than that the typical values for ground-based radars (Sherman, 2022).

Definition 1: *Signal-to-clutter ratio* (SCR) is the ratio of the power of the desired signal to the power of the clutter, typically expressed in decibels.

Note: The term "clutter" originates in the radar literature. A radar receives returns from many sources. Besides reflections from the target(s), the returns include backscatter from the environment and non-targets; these undesired returns are called "clutter" (Greco et al., 2014).

An HGV may not only appear dim but also small in infrared images, to the degree that it may occupy 0.1% or less of the image area (Dai et al., 2021). Methods for detecting dim and small targets in images can be found in the research area of *small target detection*.

3.1 Small target detection

HGVs not only appear small, with a low SCR, in infrared images, but also move at high speed. A fast-changing image background rules out efficient sequential detection algorithms such as matched filtering (which dates back to the 1980s) and necessitates detection to be performed on every image frame. Among the algorithms that use handcrafted features, Chen et al.'s (2014) local contrast method (LCM) and its derivatives were popular. Among the algorithms that use deep features, the first batch of results are widely attributed to Liu et al. (2018), who used a five-layer perceptron.

In recent years, algorithms based on the U-Net architecture (Ronneberger et al., 2015) have gained dominance. the U-Net architecture consists of contracting/down-sampling/encoder layers connected via a bottleneck for high-frequency noise filtering to expansive/up-sampling/decoder layers. Initially proposed for medical image segmentation, the architecture is now widely used in remote sensing, including small target detection.

The deep residual network (ResNet, see He et al., 2016) and feature pyramid network (Lin et al., 2017) are widely used as building blocks. ResNet uses shortcut/skip connection (i.e., connections between network layers that skip one or more layers) with identity mapping to solve the problem of vanishing/exploding gradients. Feature

pyramid networks provide features at multiple scales, so that low-level features containing fine details of the target as well as high-level edge/shape features can be simultaneously extracted.

Cross-layer feature fusion (CLFF) refers to the combination of features at different levels for context extraction. This is important for highlighting the features of dim and small targets in deep layers and subsequent accurate segmentation of the target, and is done differently in different detection algorithms. Table 1 compares some recently proposed deep neural networks. Table 1 excludes solutions that analyse multiple frames at a time because they do not work in real time, and solutions (e.g., Wu et al., 2022) not evaluated on a public dataset.

Table 1: Comparing select deep neural networks for small target detection. Datasets SIRST and IRSTD-1k are from Dai et al. (2021) and Zhang et al. (2022) respectively. P_d = % probability of detection. F_a = % false alarm rate. nIoU = % normalised intersection over union.

Reference	Name and main components	Notable original results
Dai et al., 2021	ALCNet: ResNet-20 as backbone in a feature pyramid network architecture, that uses local contrast to highlight the features of the target in high-level layers, and bottom-up local attention modulation for CLFF.	SIRST: nIoU = 72.8.
Tong et al., 2021	EAAU-Net: Like ALCNet except bottom-up attention modulation is asymmetric and uses (i) fully connected layers instead of point-wise convolution layers, and (ii) global averaging pooling for aggregating global contextual information. Furthermore, channel shuffle fusion of channel attention and spatial attention enables information flow between feature groups.	SIRST: nIoU = 74.5.
Zuo et al., 2022	AFFPN: ResNet-20 as the backbone in a feature pyramid network architecture, that uses atrous/dilated spatial pyramid pooling for CLFF, and attention fusion to enable information flow between feature groups.	SIRST: nIoU = 75.9.
Zhang et al., 2022	ISNet: U-Net network incorporating (i) edge features extraction by solving second-order finite difference equations, (ii) bottleneck convolution for low-pass filtering, (iii) deformable convolution across rows and columns for CLFF.	SIRST: $P_d = 99.2$, $F_a = 4.9$, nIoU = 78.1. IRSTD-1k: $P_d = 95.6$, $F_a = 15.4$, nIoU = 64.8.

Comparing computational complexities has not been a popular practice, but using Dai et al.'s (2021) open-source ALCNet code, we found 20-25 frames per second of processing rate is achievable on an NVIDIA RTX 3060 GPU. This is consistent with Zuo et al.'s (2022) report. Discounting the figure by an order of magnitude, small target detection should be realisable on space-borne IR sensors at several frames per second.

Once a target is detected, tracking can be done by repeating detection or by well-established object tracking techniques in the field of computer vision. Note the problem of IR-based tracking by space-borne sensors is distinctly different from IR-based tracking by hypersonic missile interceptors, which appears to have been accomplished in the 1990s (Oron et al., 1997).

3.2 Ambiguities and knowledge gaps

Ultraviolet (UV) emission from the shock layer of an HGV in near space present distinct spectral features in the 200-300 nm region when measured from any angle (Niu et al., 2016). This suggests multispectral IR/UV sensing provides a more robust detection mechanism than using IR alone (Labbé and Ghanmi, 2022), but multispectral small target detection work has yet to be done.

Dai et al. (2021) supplied code, which we have used to reproduce their accuracy results for ALCNet to within 5% of their original figures, but most other authors have not helped with this aspect of reproducible research. Fortunately, sample code for most of the deep learning building blocks can be found in the public domain, and the soundness of the algorithms proposed to date suggests enough is known about the algorithmic building blocks of an IR sensing-based HGV detection algorithm.

Dai et al. (2021) and Zhang et al. (2022) supplied training datasets, but neither appears to contain images of a fast-moving target. It seems unlikely a public dataset would contain images of an HGV, but this does not rule out the need for images of a fast-moving, dim and small target for model validation.

Given the important yet sensitive nature of an HGV dataset, an underexplored aspect is the *backdoor attack*. Classified by MITRE under the attack technique of "backdoor ML model" (MITRE, 2022), a backdoor attack attempts to embed a stealthy and persistent artifact into a machine learning model, such that the model performs normally on benign test samples, but produces the attacker's desired output when activated by an

attacker-implanted trigger in adversarial samples (Li et al., 2022). Backdoor attack can happen in any stage of the training process and is most commonly conducted by poisoning training samples. Poisoning attacks pose realistic and immediate threats to the development of small target detection and warrant close investigations.

4. Radar-based sensing

While EO/IR sensing provides only bearing information, radars have the desirable feature of providing range/distance information as well. Radars can be deployed both on the Earth's surface (ground, sea) and in space (Meng et al., 2020; Sherman, 2022; Bian et al., 2020).

Radar works by (i) bouncing modulated radiofrequency (RF) electromagnetic waves off a potential target, (ii) collecting through a high-gain RF receiver the reflected energy (called "echo") backscattered by the target in the same chosen direction, and (iii) extracting from the received signals information such as range and velocity of the target.

Compared to thermal sensing, electromagnetic sensing is not any less challenging, because an HGV flying at hypersonic speed in near space has ionised gas molecules (i.e., plasma) enveloping the vehicle – this is called a "plasma sheath". This inhomogeneous plasma sheath absorbs, reflects, and scatters electromagnetic waves, resulting in two immediate problems: (i) communication waves from within the HGV experiences blackout, (ii) scattered radar echoes become attenuated and irregular (Bian et al., 2020; Ma et al., 2021). The latter is relevant here as the backscattered radar cross section (RCS, see Definition 2) of an HGV becomes hard to predict. The relations between each of these variables – radar frequency, radar altitude, HGV altitude, HGV speed – and dynamic RCS are discernible (Bian et al., 2020), but knowledge of these relations has yet to lead to a robust HGV detection scheme.

Definition 2: The backscattered / monostatic / target radar cross section (RCS, see Mahafza and Elsherbeni, 2004) of a target R distance away from a radar is

$$\sigma = 4\pi R^2 \lim_{R\to\infty} \left(P_{Dr}(R)/P_{Di}(R)\right),$$

where P_{Dr} is the power density of the scattered waves at the receiving antenna, and P_{Di} is the power density of the waves incident on the target. Dynamic RCS is an extension of backscattered RCS that accounts for relative motion between the radar and target.

The plasma frequency lies in the range of either 0.1-100 GHz (Ma et al., 2021) or 0.284-28 GHz (Musselman and Chastain, 2020). Two solutions have been proposed:

1. Increase radar frequency to THz or at least the Ka band (26-40 GHz) to improve radar penetration of the plasma sheath. However, THz technologies are still in infancy and Ka-band radars are close-range radars.
2. Decrease radar frequency to the high frequency (HF) band (3-30 MHz) to detect the plasma sheath of an HGV instead. Beyond-line-of-sight detection becomes possible as skywave over-the-horizon radars bounce HF-band signals off the ionosphere to potential targets and collects reflections back from the targets through the same propagation path.

Interestingly, inverse synthetic aperture radar remains effective in the X band (8-12 GHz, see Xie et al., 2022). For circumventing the plasma sheath problem, the common radar bands of L, C and Ku are not as useful.

In terms of real-world implementation, most existing space observation radars have a low resolution (Bian et al., 2020). Pulse compression is commonly assumed for the coherent processing (signal-to-noise ratio maximisation) of echo signals (Xu and Liao, 2014), but pulse compression can be challenging to implement due to the high bandwidth requirement. Long-range radars like the Long-Range Discrimination Radar (MDA, 2022) are too bulky, heavy, and energy-intensive to be space-borne.

4.1 Ambiguities and knowledge gaps

In the literature, there is little mention of concrete plans for space-bound deployment of long-range radars. Both the plasma sheath problem and space-based radar technologies present challenges and uncertainties. Assuming terrestrial long-range radars and over-the-horizon radars are the way forward, data fusion across space-based and ground-based sensing regimes warrants concrete exploration.

5. Aerodynamics modelling

Following the justification in Sec. 2, this section discusses physics-based mathematical modelling of hypersonic aerodynamics. Most HGV tracking algorithms in the literature are based on different variations of the aerodynamic model in Eq. (**Error! Reference source not found.**), expressed in the velocity-turn-climb (VTC) coordinate system (also known as semi-speed / half-velocity / trajectory / flight-path coordinate system).

$$\dot{r} = v \sin\theta, \tag{1a}$$

$$\dot{\theta} = \frac{v \cos\gamma \sin\psi}{r \cos\phi}, \tag{1b}$$

$$\dot{\phi} = \frac{v \cos\gamma \cos\psi}{r}, \tag{1c}$$

$$\dot{v} = -a_D - g \sin\gamma, \tag{1d}$$

$$\dot{\gamma} = \frac{1}{v}\left\{a_L \cos\sigma + \left(\frac{v^2}{r} - g\right)\cos\gamma\right\}, \tag{1e}$$

$$\dot{\psi} = \frac{1}{v}\left\{\frac{a_L \sin\sigma}{\cos\gamma} + \frac{v^2}{r}\cos\gamma \sin\psi \tan\phi\right\}. \tag{1f}$$

In Eq. (**Error! Reference source not found.**), r is the radial distance of the HGV from the Earth's centre; θ and ϕ measure the longitude and latitude of the HGV; v is the HGV's speed; γ is the flight-path angle; ψ is the velocity heading angle; σ is the bank angle; g is the gravitational acceleration; a_D and a_L are the magnitudes of the drag and lift accelerations.

Eq. (**Error! Reference source not found.**) is based on the prevalent simplifications (Li et al., 2015) that (i) the Earth is a nonrotating uniform sphere, and (ii) the HGV's sideslip angle is negligible, leaving attitude and trajectory to be fully described by the angle of attack and bank angle. Furthermore, the commanded/desired acceleration is assumed to be zero (Hough, 2017; Cheng et al., 2020).

Most academic researchers based their model on the Hypersonic Technology Vehicle 2 (HTV-2), a research vehicle in the now completed DARPA Falcon program (Walker et al., 2008), and Common Aero Vehicle (CAV, see Phillips, 2003). We note HGV models are dominated by reentry dynamics (Hough, 2017), and cannot be based on the widely cited hypersonic vehicle models by Shaughnessy et al. (1990) and Parker et al. (2007). However, the absence of flight test data serving as ground truth poses two difficulties: (i) parameter values cannot be estimated from data, (ii) models cannot be validated. Academic researchers have been coping with these difficulties through *parametric modelling*, where a geometric model (Cheng et al., 2020) or aerodynamic model (Liu et al., 2018b) or manoeuvre model (Hu et al., 2021) is parameterised and subjected to simulations for the determination of suitable parameter values. For example,

- Li et al. (2019) applied the oblique shock and Prandtl-Meyer expansion flow theory to express accelerations in VTC coordinates as functions of three parameters: angle of attack, bank angle and mass (the first two of which are in fact variables). The values of these parameters were constrained to fixed ranges, which were in turn discretised into 12 value points, resulting in a total of 12×12×12 possible models. Out of these many possible models, 12 models were selected as the model-set for an interacting multiple model (IMM) filter (see Sec. 6).
- Cheng et al. (2020) sampled the geometrical design space of HTV-2 to determine five possible geometrical configurations. For each of the five configurations, computational fluid dynamics simulations were (likely) used to calculate the values of three parameters, namely lift coefficients, drag coefficients and lift-to-drag ratio, for different angles of attack and Mach numbers.

To generate/simulate HGV trajectories, besides an aerodynamic model, representative values of guidance commands are also needed. Assuming acceleration command is zero (as mentioned earlier), the commanded angle of attack and bank angle still need to be specified based on type of manoeuvre. Different classes of manoeuvres have emerged in the literature, but Li et al.'s (2015) classification seems to cover most cases: (i) equilibrium glide, (ii) longitudinal skip glide, (iii) lateral turning manoeuvre, (iv) lateral weaving manoeuvre.

5.1 Ambiguities and knowledge gaps

Seeing a significant amount of work published in generic open-access journals rather than aeronautics journals raises the question of whether there has been enough expert scrutiny. Even where the methodology is sound, there are difficulties in reproducing the published models. For example,

- In (Feng et al., 2017), lift and drag accelerations should not depend on the Earth's radius.
- In (Li et al., 2019), the analytical expressions for the accelerations in VTC coordinates are missing.
- In (Cheng et al., 2020), some symbols are undefined and multiple parameter values are missing, including three time constants; and the lift and drag coefficients as functions of angle of attack and Mach number.

Missing parameter values and model details hamper reproducible research.

6. HGV tracking

This section discusses two main tracking paradigms: stochastic estimation and machine learning.

Stochastic estimation: Target tracking or trajectory estimation using a state estimation filter is a well-established field, but the hypersonic context is new. The distinct manoeuvre modes of an HGV make the IMM filter a natural algorithmic component and thus feature prominently in Table 2. Distributed bearing-only filtering schemes are a good fit for the NDSA (see Sec. 2). Wang and Li's (2021) in Table 2 is an example of such a scheme.

Table 2: **Comparing select filtering schemes for tracking HGVs.**

Reference	Tracking filter	Notable aspects
Fan et al., 2017	Hybrid grid multiple model and Cubature Kalman filters	Uses multiple models in the initial tracking stage but single model in the stable tracking stage.
Feng et al., 2017	IMM with updates to Markov transition matrix	Trajectories not based on any aerodynamic model. Instead, acceleration is modelled as a zero-mean stochastic process with sinusoidal autocorrelation.
Li et al., 2019	IMM and unscented Kalman filters	Using 6 models in the model-set achieves the same level of accuracy as using 12.
Cheng et al., 2020	Iterative extended Kalman filter	As accurate as but more efficient than unscented Kalman filter.
Wang and Li, 2021	Distributed information-weighted consensus filtering incorporating cubature point calculation rules from cubature Kalman filter	Can work with IR sensors because it uses only bearing measurements, but requires multiple sensors. Cubature rules facilitate efficient nonlinear estimation.

Machine learning: Viewed as a time series, a trajectory can be estimated using multivariate time series forecasting methods. Due to their feature extraction capabilities, recurrent neural networks (with or without gated recurrent units), long short-term memory (LSTM), and transformer models have become popular methods. Common strategies include the following:

- Trajectory prediction is divided into two parts: linear and nonlinear. For example, Xie et al. (2021) applied a recurrent neural network with bidirectional gated recurrent units to the nonlinear part, and a single-layer perceptron to the linear part. Zhang et al. (2022) applied the seq2seq architecture to the nonlinear part and also a single-layer perceptron to the linear part.
- Use the convolution operator to extract spatial features and LSTM for extracting temporal features; see (Zeng et al., 2021) and (Zhang et al., 2022).

7. Ambiguities and knowledge gaps

In general, papers that focussed on tracking used a simplified model, but without ground-truth data, the impact of these simplifications is unclear. Following Wang and Li (2021), the natural next step is distributed tracking using bearing-only measurements, that is resilient to false data injection attacks, where rogue sensors feed false data to its network neighbours. The relevant research lies in the area of secure/resilient distributed estimation/filtering, which is a maturing area, but not in the hypersonic context.

For the evaluation of tracking filters, values for process noise, measurement noise and the initial error covariance matrix – when they are provided at all – generally require more rigorous justification and investigation. For example, these values are missing in (Cheng et al., 2020) and not justified in (Wang and Li, 2021).

For the evaluation of machine learning methods, absence of ground-truth data means whichever model is used to generate the training and test trajectories is taken for granted. Radar measurements should be modelled.

Since stochastic estimation is a mathematically rigorous and more established field, comparing machine learning methods with tracking filters serve clear scientific interests.

8. Conclusion and future directions

The intrinsic ability of the Tracking Layer (see Sec. 2) to maximise *precision* and *recall* is clearly important, but our main concern here is security, as a disruption to the kill chain can have disastrous consequences. More than security, *resilience* accounts for the inevitability of an attack by detecting, containing and resolving any security breach (Thompson et al., 2016).

Resilience requires intercepting insider adversaries who understand and can infiltrate the system. Hence, defenders also need to understand the system beyond its ICT functions, including the information flows across cyber-physical boundaries (e.g., sensing and tracking). Motivated by this need, this paper presents a highly compressed review and analysis of the latest advances in HGV detection and tracking.

Based on our review, we conclude that (i) small target detection is sound and promising, (ii) bearing-only multispectral sensing necessitates distributed detection and tracking by a satellite constellation, (iii) data fusion across space-borne multispectral sensors and terrestrial radars is necessary. Uncertainties encountered include: (i) the applicability of space-borne radars; (ii) the adequacy of existing hypersonic flight models for tracking scheme development; (iii) the merits of machine learning methods over estimation filters for tracking.

The preceding observation explains the priorities of the NDSA, and motivates as future work (i) a threat assessment of the Tracking Layer assuming the existence of aforementioned components and accordingly (ii) a security plan. A fertile ground for cybersecurity research is the feasibility of backdoor learning (see Sec. 3.2), i.e., the feasibility of releasing a rigged dataset so that small target detection algorithms trained on it malfunction in pre-planned scenarios. Another promising direction is attack-resilient distributed tracking using bearing-only measurements (see Sec. 7).

Acknowledgement

Dr Sanjeev Arulampalam, Dr Ronald Mulinde and Timothy Pitts from DSTG provided valuable input.

References

Bian, Z., Li, J. and Guo, L. (2020) 'Simulation and Feature Extraction of the Dynamic Electromagnetic Scattering of a Hypersonic Vehicle Covered with Plasma Sheath', *Remote Sensing*, vol. 12, no. 17. DOI: 10.3390/rs12172740.

Chen, C. L. P., Li, H., Wei, Y., et al. (2014) 'A Local Contrast Method for Small Infrared Target Detection', *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581. DOI: 10.1109/TGRS.2013.2242477.

Cheng, Y., Tang, S., Lyu, S., et al. (2020) 'Aerodynamic Statistics-Based Trajectory Estimation of Hypersonic Maneuvering Target', *IEEE Access*, vol. 8, pp. 227642–227656. DOI: 10.1109/ACCESS.2020.3044589.

Dai, Y., Wu, Y., Zhou, F. and Barnard, K. (2021) 'Attentional Local Contrast Networks for Infrared Small Target Detection', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824. DOI: 10.1109/TGRS.2020.3044958. Dataset at https://github.com/YimianDai/sirst.

Fan, Y., Lu, F., Zhu, W., et al. (2017) 'A Hybrid Model Algorithm for Hypersonic Glide Vehicle Maneuver Tracking Based on the Aerodynamic Model', *Applied Sciences*, vol. 7, no. 2. DOI: 10.3390/app7020159.

Feng, S., Tan, Q., Li, Q. and Ren, Z. (2017) 'Trajectory tracking for hypersonic glide vehicles based on improved Sine-AIMM', *CCC*, pp. 5475–5480. DOI: 10.23919/ChiCC.2017.8028225.

Greco, M. S. and Watts, S. (2014) 'Chapter 11 - Radar Clutter Modeling and Analysis', in *Academic Press Library in Signal Processing: Volume 2*, Elsevier, vol. 2, pp. 513–594. DOI: 10.1016/B978-0-12-396500-4.00011-9.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep Residual Learning for Image Recognition', *CVPR*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hough, M. E. (2017) 'Reentry Maneuver Estimation Using Nonlinear Markov Acceleration Models', *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 7, pp. 1693–1710. DOI: 10.2514/1.G002577.

Hu, Y., Gao, C., Li, J. and Jing, W. (2021) 'Maneuver mode analysis and parametric modeling for hypersonic glide vehicles', *Aerospace Science and Technology*, vol. 119, p. 107166. DOI: 10.1016/j.ast.2021.107166.

Labbé, P. and Ghanmi, A. (2022) *Current and future hypersonic threats, scenarios and defence technologies for the security of Canada*, Defence Research and Development Canada Scientific Report. Available at https://cradpdf.drdc-rddc.gc.ca/PDFS/unc385/p814591_A1b.pdf.

Li, G., Zhang, H. and Tang, G. (2015) 'Maneuver characteristics analysis for hypersonic glide vehicles', *Aerospace Science and Technology*, vol. 43, pp. 321–328. DOI: https://doi.org/10.1016/j.ast.2015.03.016.

Li, S., Lei, H., Shao, L. and Xiao, C. (2019) 'Multiple Model Tracking for Hypersonic Gliding Vehicles With Aerodynamic Modeling and Analysis', *IEEE Access*, vol. 7, pp. 28011–28018. DOI: 10.1109/ACCESS.2019.2899678.

Li, Y., Jiang, Y., Li, Z. and Xia, S.-T. (2022) 'Backdoor Learning: A Survey', *IEEE Transactions on Neural Networks and Learning Systems*, early access. DOI: 10.1109/TNNLS.2022.3182979.

Lin, T.-Y., Dollár, P., Girshick, R., et al. (2017) 'Feature Pyramid Networks for Object Detection', *CVPR*, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

Liu, M., Du, H.-Y., Zhao, Y.-J., et al. (2018a) 'Image Small Target Detection based on Deep Learning with SNR Controlled Sample Generation', *Current Trends in Computer Science and Mechanical Automation Vol.1: Selected Papers from CSMA2016*, De Gruyter, pp. 211–220. DOI: 10.1515/9783110584974-025.

Liu, Y., Chen, B., Li, Y. and Shen, H. (2018b) 'Overview of control-centric integrated design for hypersonic vehicles', *Astrodynamics*, vol. 2, no. 4, pp. 307–324. DOI: 10.1007/s42064-018-0027-8.

Mahafza, B. R. and Elsherbeni, A. Z. (2004) *MATLAB Simulations for Radar Systems Design*, CRC Press.

Ma, Z., Zeng, Y., Deng, B., et al. (2021) 'Echo attenuation characteristics of a vehicle covered with the inhomogeneous plasma sheath in terahertz band', *Tenth International Symposium on Ultrafast Phenomena and Terahertz Waves (ISUPTW 2021)*, SPIE, vol. 11909, p. 119090F. DOI: 10.1117/12.2604446.

MDA (2022) *Long Range Discrimination Radar (LRDR), Clear Space Force Station (CSFS), Alaska*, Fact Sheet [Online]. Available at https://www.mda.mil/global/documents/pdf/lrdr.pdf.

Meng, F. and Tian, K. (2020) 'Phased-Array Radar Task Scheduling Method for Hypersonic-Glide Vehicles', *IEEE Access*, vol. 8, pp. 221288–221298. DOI: 10.1109/ACCESS.2020.3043338.

MITRE (2022) Backdoor ML Model [Online]. MITRE ATLAS Technique ID AML.T0018. Available at https://atlas.mitre.org/techniques/AML.T0018.

Munoz, C. (2021) Next-Gen OPIR sensor payload clears critical milestone [Online]. Jane's International Defence Review. Available at https://customer.janes.com/Janes/Display/BSP_4061-IDR.

Musselman, R. L. and Chastain, S. H. (2020) 'Beyond LOS Detection of Hypersonic Vehicles', *2020 International Applied Computational Electromagnetics Society Symposium (ACES)*, pp. 1–2. DOI: 10.23919/ACES49320.2020.9196080.

Niu, Q., He, Z. and Dong, S. (2016) 'Prediction of shock-layer ultraviolet radiation for hypersonic vehicles in near space', *Chinese Journal of Aeronautics*, vol. 29, no. 5, pp. 1367–1377. DOI: 10.1016/j.cja.2016.04.021.

Oron, E., Bar-Shalom, Y. and Lachish, M. (1997) 'Advanced IR imaging and tracking for hypersonic intercept', *1997 IEEE Aerospace Conference*, vol. 1, pp. 139–148. DOI: 10.1109/AERO.1997.574403.

Parker, J. T., Serrani, A., Yurkovich, S., et al. (2007) 'Control-Oriented Modeling of an Air-Breathing Hypersonic Vehicle', *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 3, pp. 856–869. DOI: 10.2514/1.27830.

Phillips, T. H. (2003) A Common Aero Vehicle (CAV) Model, Description, and Employment Guide, Schafer Corporation for AFRL and AFSPC. Only available from third-party websites.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

SDA (2021) 'SDA About Us 2020 to 2021: Speed. Delivery. Agility', [Online]. Available at https://www.sda.mil/wp-content/uploads/2022/01/SDA-About-Us-2020-to-2021.pdf.

Shaughnessy, J. D., Pinckney, S. Z., Mcminn, J. D., et al. (1990) *Hypersonic Vehicle Simulation Model: Winged-Cone Configuration*, NASA Technical Memorandum 102610. Available at https://ntrs.nasa.gov/citations/19910003392.

Sherman, J. (2022) 'Hypersonic Weapons Can't Hide from New Eyes in Space', *Scientific American* [Online]. Available at https://www.scientificamerican.com/article/hypersonic-weapons-cant-hide-from-new-eyes-in-space/.

Singer, R. A. (1970) 'Estimating Optimal Tracking Filter Performance for Manned Maneuvering Targets', *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-6, no. 4, pp. 473–483. DOI: 10.1109/TAES.1970.310128.

Strout, N. (2019) 'The small sat solution to hypersonic weapons, explained', [Online]. Available at https://www.c4isrnet.com/battlefield-tech/2019/08/06/the-small-sat-solution-to-hypersonic-weapons-explained/.

Thompson, M. A., Ryan, M. J., Slay, J. and McLucas, A. C. (2016) 'Harmonized taxonomies for security and resilience', *Information Security Journal: A Global Perspective*, Taylor & Francis, vol. 25, no. 1–3, pp. 54–67. DOI: 10.1080/19393555.2016.1145772.

Tong, X., Sun, B., Wei, J., et al. (2021) 'EAAU-Net: Enhanced Asymmetric Attention U-Net for Infrared Small Target Detection', *Remote Sensing*, vol. 13, no. 16. DOI: 10.3390/rs13163200.

Tracy, C. L. and Wright, D. (2020) 'Modeling the Performance of Hypersonic Boost-Glide Missiles', *Science & Global Security*, Routledge, vol. 28, no. 3, pp. 135–170. DOI: 10.1080/08929882.2020.1864945.

Van Wie, D. M. (2021) 'Hypersonics: Past, Present, and Potential Future', *John Hopkins APL Technical Digest*, vol. 35, no. 4, pp. 335–341.

Walker, S., Sherk, J., Shell, D., et al. (2008) 'The DARPA/AF Falcon Program: The Hypersonic Technology Vehicle #2 (HTV-2) Flight Demonstration Phase', *15th AIAA International Space Planes and Hypersonic Systems and Technologies Conference*. DOI: 10.2514/6.2008-2539.

Wang, X. and Li, C. (2021) 'Cooperative Tracking of Hypersonic Target with Bearing-Only Measurements', *2021 40th Chinese Control Conference (CCC)*, pp. 5212–5216. DOI: 10.23919/CCC52363.2021.9549405.

Wu, D., Cao, L., Zhou, P., et al. (2022) 'Infrared Small-Target Detection Based on Radiation Characteristics with a Multimodal Feature Fusion Network', *Remote Sensing*, vol. 14, no. 15. DOI: 10.3390/rs14153570.

Xie, Y., Li, X., Shen, F., et al. (2022) 'Influence of Plasma Sheath's Velocity Field on ISAR Imaging of Hypersonic Target', *Remote Sensing*, vol. 14, no. 15. DOI: 10.3390/rs14153799.

Xie, Y., Zhuang, X., Xi, Z. and Chen, H. (2021) 'Dual-Channel and Bidirectional Neural Network for Hypersonic Glide Vehicle Trajectory Prediction', *IEEE Access*, vol. 9, pp. 92913–92924. DOI: 10.1109/ACCESS.2021.3092515.

Xu, X. and Liao, G. (2014) 'MPD model for radar echo signal of hypersonic targets', *The Journal of Engineering*, vol. 2014, no. 8, pp. 399–406. DOI: 10.1049/joe.2014.0097.

Zeng, J., Gao, Y. and Ding, D. (2019) 'Positioning and Tracking Performance Analysis of Hypersonic Vehicle based on Aerodynamic Model', *ICAICA*, pp. 312–318. DOI: 10.1109/ICAICA.2019.8873439.

Zeng, K., Zhuang, X., Xie, Y. and Xi, Z. (2021) 'Hypersonic Vehicle Trajectory Classification Using Improved CNN-LSTM Model', *ICUS*, pp. 691–696. DOI: 10.1109/ICUS52573.2021.9641452.

Zhang, J., Xiong, J., Li, L., et al. (2022) 'Motion State Recognition and Trajectory Prediction of Hypersonic Glide Vehicle Based on Deep Learning', *IEEE Access*, vol. 10, pp. 21095–21108. DOI: 10.1109/ACCESS.2022.3150830.

Zhang, M., Zhang, R., Yang, Y., et al. (2022) 'ISNet: Shape Matters for Infrared Small Target Detection', CVPR, pp. 867–876. DOI: 10.1109/CVPR52688.2022.00095. Dataset at https://github.com/RuiZhang97/ISNet.

Zhang, Q.-L. and Yang, Y.-B. (2021) 'SA-Net: Shuffle Attention for Deep Convolutional Neural Networks', *ICASSP*, pp. 2235–2239. DOI: 10.1109/ICASSP39728.2021.9414568.

Zuo, Z., Tong, X., Wei, J., et al. (2022) 'AFFPN: Attention Fusion Feature Pyramid Network for Small Infrared Target Detection', *Remote Sensing*, vol. 14, no. 14. DOI: 10.3390/rs14143412.

Cyber Security Training in Finnish Basic and General Upper Secondary Education

Martti Lehto[1] and Pekka Neittaanmäki[2]
[1]Faculty of Information Technology, University of Jyväskylä, Finland
[2]Faculty of Information Technology, University of Jyväskylä, Finland

martti.lehto@jyu.fi
pekka.neittaanmaki@jyu.fi

Abstract: Cyber security in Finland is part of other areas of comprehensive security, as digital solutions multiply in society and technologies advance. Cyber security is one of the primary national security and national defense concerns. Cyber security has quickly evolved from a technical discipline to a strategic concept. Cyber security capacity building can be measured based on the existence and number of research and developments, education and training programs, and certified professionals and public sector agencies.Cybersecurity awareness and the related civic skills play an increasingly important role as our societies become more digitalized. Improving citizens' cyber skills through education is an important goal that would strengthen Finland as a country of higher education and expertise and lay the foundation for the society of the future. Pursuant to the Finland's Cyber Security Strategy (2019) "National cyber security competence will be ensured by identifying requirements and strengthening education and research." Finland's Cyber Security Development Programme (2021) necessitates that in basic education ensures young people have sufficient skills to operate in a digital operating environment and that they understand cyber security threats and know how to protect themselves from them. So, cybersecurity is an important subject for everyone, not just industry or public organizations. It's also vital for our children to understand how to stay safe online, and the need to be aware of any dangers that might come their way. Cybersecurity awareness training is important because it teaches pupils how they can protect themselves from cyber-attacks (MTC, 2021). The study of cybersecurity education in Finland was made in autumn 2021 and spring 2022 for the National Cyber Security Director. According to the study, measures are needed so that cyber security becomes an important aspect when planning education and teaching. There are different models to choose from to make training more effective. This paper presents the results of the research focusing basic and general upper secondary education.

Keywords: Cyber Security Education, Basic Education, General Upper Secondary Education

1. Introduction

Finnish society needs cyber security competence both in public administration and in the business community. At the national level, it must ensure that companies have both top-level experts and other competent personnel. Skills development is also emphasized in the cooperation between the business community and research. (Security Committee, 2019)

Finland is a world leader in using digitalization in higher education and in continuous learning in higher education. Digitalization aims to make educational content as widely available as possible. In view of the growing needs for skills renewal, continuous learning should be given greater priority in the higher education sector. According to the Education Policy Report of the Finnish Government (VN, 2021), a higher level of competence and top expertise are required to develop Finnish society and well-being. One of the objectives is that by 2030, at least half of all young adults in Finland will complete a higher education degree.

Digital transformation has changed societies with an ever-deepening impact on everyday life and demonstrated the need for larger digital education and training in every level. The digital future at school is already integrated into the everyday lives of learners. The pandemic-related exceptional situation during Covid-19 has made it clear to us how important the digitization of state institutions is. At home, smartphones have become indispensable for students. During digitization, schools will transform analog, location-based teaching methods into digital, location-independent training opportunities. The digitization of the schools requires cyber secure systems and cyber security skills from teachers and pupils.

Digital environment has security risks for young people. They are among others social withdrawal, digital gaming addiction, cyberbullying, cognitive overload, material that is harmful for the age, a decrease in mental balance, fake people.

The study, commissioned by the Ministry of Transport and Communications from the University of Jyväskylä leaded consortium, examined the development needs in cybersecurity education and explored comprehensively the quantitative and qualitative development of cybersecurity competencies. This paper is part of this larger study and gives recommendations. (Lehto, 2022)

2. The Finnish Education System

The Finnish education system is a mixture of state controlled or steered and relatively autonomous elements. The government determines the general objectives of education and the division of classroom hours between different subjects. The Ministry of Education and Culture (MEC, 2022) drafts legislation and government decisions pertaining to education. The Finnish National Agency for Education lays out the concrete objectives and core contents of instruction in the different subjects and is responsible for the national core curriculum with its directive norms for good achievement in each. Local authorities (generally municipalities) are responsible for the practical arrangement of schooling and for composing the municipal curriculum based on the national core curriculum. Each school, in turn, writes its own curriculum based on both the national core curriculum and the municipal document. The Finnish education system consists of (FNAE, 2022):

- Early childhood education and care which is provided for children before the compulsory education begins,
- Pre-primary education which is provided for children in the year preceding the beginning of compulsory education,
- Nine-year basic education (comprehensive school), which is compulsory,
- Upper secondary education, which is either general upper secondary education or vocational education and training, and
- Higher education provided by universities and universities of applied sciences.
- Adult education is available at all levels.

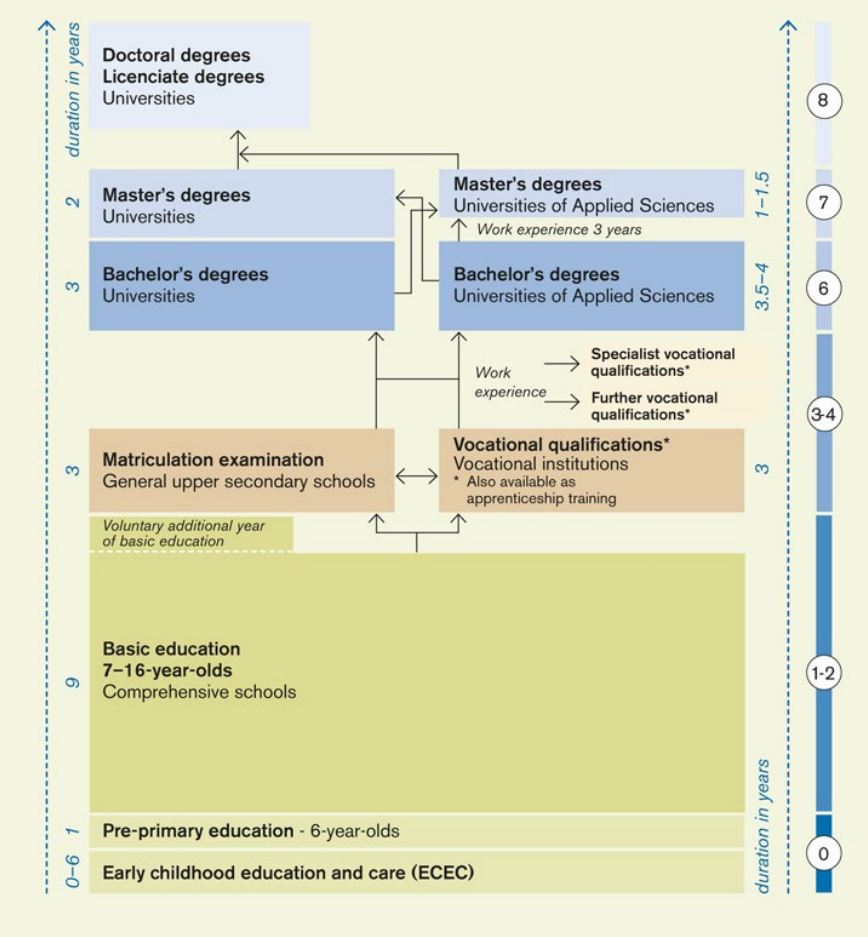


Figure 1: illustrates the Finnish education system.

3. Basic education

3.1 Fundamentals of basic education

Basic education starts in the year when a child turns seven and lasts nine years. basic education starts with comprehensive school and ends at the age of 18. Comprehensive school education (basic education) consists of school years 1 to 9 and is meant for all children aged between 7 and 17 (whole age group). All children who reside permanently in Finland must attend compulsory education. Basic education is provided within a single structure, that is, there is no division into primary and lower secondary education. Instruction is usually given by the same class teacher in most subjects in the first six year-classes and by subject specialists in the last three years. In the curriculum, transversal competence refers to a combination of knowledge, skills, values, attitudes, and will. Competence also means the ability to apply knowledge and skills in a given situation. (FNAE, 2014.) The transversal competence areas are:

1. Thinking and learning to learn
2. Cultural competence, interaction, and self-expression
3. Taking care of oneself and managing daily life
4. Multiliteracy
5. ICT competence
6. Working life competence and entrepreneurship
7. Participation, involvement, and building a sustainable future

Cybersecurity and information security are not a separate subject in the curriculum. Instead, the curriculum includes information and communication technologies (ICT) as a transversal subject. This means that transversal learning outcomes are integrated into the intended learning outcomes of different subjects in different grades. The curriculum states that "ICT is methodically exploited in all grades of basic education, in different subjects and multidisciplinary learning modules, and in other schoolwork." The organiser of the teaching is responsible for planning the content of the teaching in accordance with the curriculum. (FNAE, 2014

One of the transversal competence objectives is ICT competence, which is an important civic skill both in itself and as part of multiliteracy. The pupils are supported in familiarising themselves with various ICT applications and uses and in observing their significance in their daily life. The aim is that they also learn to perceive its risks in a global world. ICT competence is developed in four main areas (FNAE, 2014):

1. The pupils are guided in understanding the principle of using ICT, its operating principles and key concepts and supported to develop their practical ICT competence in producing their own work.
2. The pupils are guided in using ICT responsibly, safely, and ergonomically.
3. The pupils are guided in using ICT in information management and in exploratory and creative work.
4. The pupils gather experience of and practice using ICT in interaction and networking.

The list below shows by grade the third main area, which focuses on the safe and responsible use of ICT.

- In grades 1 and 2, the aim is to discuss together with the pupils and search for safe ways to use ICT and the related etiquette.

- In grades 3 to 6, the pupils are guided in responsible and safe use of ICT, good manners, and knowledge of basic copyright principles. In their schoolwork, they practice using various communication systems and educational social media services.

- When moving to grades 7 to 9, information security and the related risks are addressed in a more concrete way: the pupils are guided to use ICT in a way that is safe and ethically sustainable. They learn how to protect themselves from possible information security risks and how to avoid losing data. They are guided towards responsible activities by reflecting on, for example, the meaning of the concepts of information protection and copyrights and the potential repercussions of irresponsible and illegal activities. (FNAE, 2014.)

3.2 Analysis of the research

Eleven towns across Finland were selected for the survey, which was sent to a total of 448 school principals. The schools were primary or lower secondary schools that follow the Finnish national curriculum for basic education. A total of 108 responses were received.

Inclusion of cybersecurity education in teachers' subjects

Teachers were asked to assess on a scale of 1 to 5 (completely disagree to fully agree) whether they include cybersecurity education as allowed by the subject requirements. The average score was 3.45. Teachers whose answers ranged between 4 and 5 represent equally both primary and lower secondary schools. However, teachers whose answers ranged between 1 and 2 mostly represent teachers in grades 7 to 9. It can be noted that in primary school, cybersecurity / information security is addressed more extensively in different subjects than in secondary school. The result is shown in Figure 2.

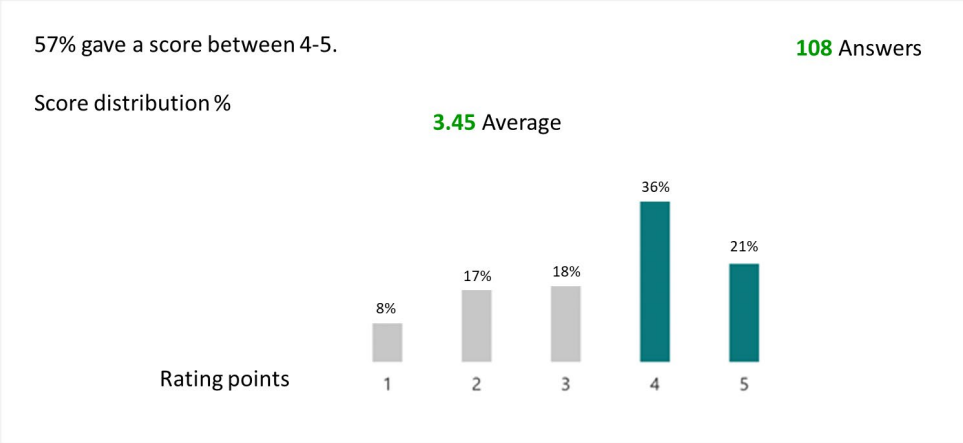


Figure 2: Inclusion of cybersecurity education in teachers' own subject.

Inclusion of cybersecurity education in other subjects

The teachers were asked whether cybersecurity education should in some way be included in all subjects or whether it should be taught separately, for example, only as part of IT education. The teachers were not unanimous on this point: 27% of the respondents felt that it would be more sensible to only teach cybersecurity as part of IT education, while 73% of the respondents felt that cybersecurity belongs to all areas of basic education.

Pupils' and teachers' awareness of secure use of services

The survey asked teachers to assess pupils' and staff's knowledge of the secure use of equipment and study-related services. The assessment was carried out on a scale of 1 to 5 (fully disagree to fully agree). The average response was 3.26. Most responses ranged between 3 and 4, suggesting that teachers perceive both their own and their pupils' knowledge of the secure use of equipment and study-related services as quite strong.

Implementation of secure use of information and communication technologies in teaching

Teachers were asked to assess whether the objectives for the safe use of information and communication technology defined in the curriculum for different grades are met in their teaching. The majority of teachers felt that the objectives were met in their teaching. Eleven teachers estimated that the grade-specific objectives are not achieved in their teaching. Of these teachers, all teach grades 7 to 9 (see Figure 3).

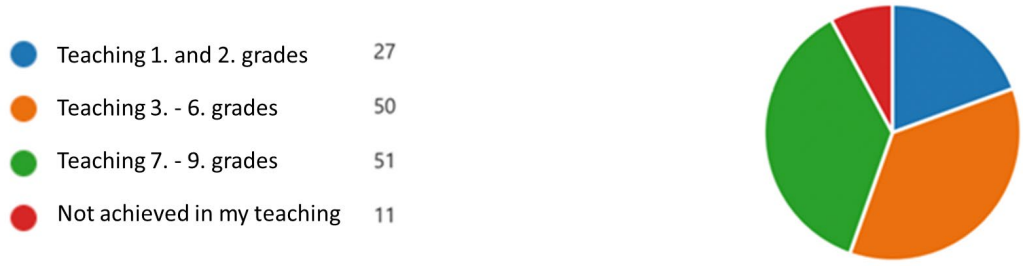


Figure 3: Objectives of secure use ICT in the curriculum by grade.

Clarity of the current curriculum guidelines for teaching cybersecurity/information security

Teachers were asked to assess on a scale of 1 to 5 (fully disagree to fully agree) how clear the instructions in the current curriculum are in support of cybersecurity and information security education. The average was 2.96 (see Figure 4). Most respondents answered 3, which indicates that the teachers receive some support from the curriculum, but do not consider the curriculum guidelines sufficiently clear when it comes to teaching cybersecurity / information security.

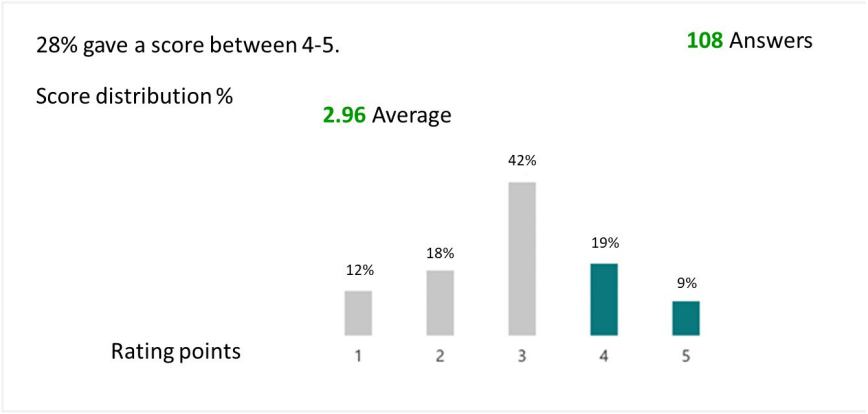


Figure 4: Curriculum guidelines for cybersecurity teaching.

Teachers' perception of their cybersecurity teaching competence

Teachers took a rather positive view of their own knowledge and ability to incorporate cybersecurity as part of their teaching. The assessment was carried out on a scale of 1 to 5 (fully disagree to fully agree). The average score was 3.66. Teachers who responded on a scale of 4 to 5 (n = 68) are fairly evenly divided between teachers of grades 3 to 6 and of grades 7 to 9. Of those responding between 1 and 3 (n = 40), a clear majority (n = 27) teach grades 7 to 9. This is in line with the answers to previous questions, where lower secondary school teachers see cybersecurity/information security as a separate subject rather than a topic to be included in all subjects. Figure 5 presents the teachers' view of their ability to teach cybersecurity.

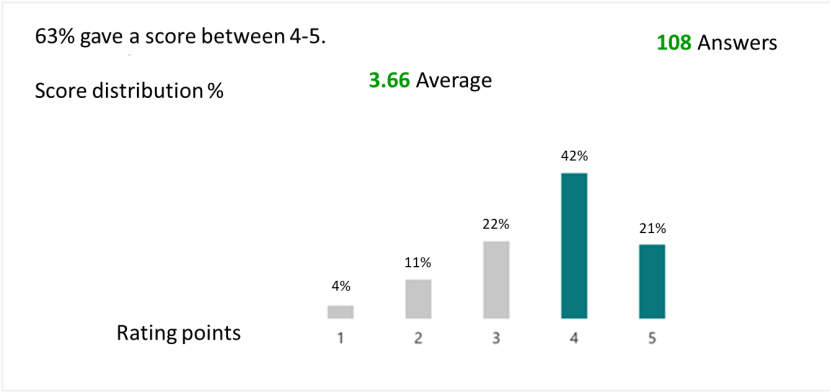


Figure 5: Teachers' self-assessment of their competence.

3.3 Development of digital safety education in basic education

3.3.1 General development needs

The training needs of digital security and its importance are widely recognised. There is thus a willingness to develop training. Recent projects, such as the *Cyber Security Development Programme*, *New Literacy Development Programme,* and the Finnish National Board of Education's guidelines for schools on information security, demonstrate a desire to make digital security an important area in the planning of education and teaching at the level of basic education.

Currently, however, materials and tools serve more as guidelines and support than as obligations, so that responsibility for the use of these materials, or lack thereof, rests with the organiser of the teaching. In addition, the objectives for the use of ICT set out in the curriculum within different subjects remain broad, so that implementation methods can vary greatly, for example, between different towns, not to mention within different schools within a town. Information and communication technologies are offered as an elective subject in many towns, but the teaching includes very little digital security. The objectives are often very similar, regardless of the grade.

It should be noted that one of the objectives of the *Cybersecurity Development Plan* is that digital security should be included in the curriculum as a separate subject. This objective would be particularly important if reached. In addition to ensuring that primary and lower secondary school pupils have digital security skills, there is also a need to ensure an adequate supply and level of education for teachers. This will cover the needs of the entire comprehensive school regarding digital security.

In order to fulfil the goal of Finland's cybersecurity strategy and to enable everyone to operate safely in the digital world, more studies must be carried out on the digital security learning needs of children in this age group.

3.3.2 Development models based on research

There are two ways of approaching the development of digital security education in primary education: the study examined teachers' views on whether digital security should be integrated into every subject or whether it should be separated into, for example, information technology education. For a number of subjects, the curriculum for basic education identifies skills that are also needed for secure behaviour in the digital world, such as source criticism and media literacy, but there is no clear stance on digital security/cybersecurity/information security directly.

The results of the survey demonstrate that the inclusion of digital security teaching in primary and lower secondary schools within the subject limits varies between teachers: some do not include digital security in their teaching at all. In addition, teachers' views on the guidelines for the development of digital security education are divided. The study highlighted three models of how digital security education can be developed and increased in basic education. The models are not mutually exclusive.

Model 1: Digital security as a concept of transversal competence

This line of development could be achieved on the basis of the current curriculum by modifying the concept of transversal competences as regards ICT skills. Currently, transversal competence consists of six areas, and it forms the common objectives of all subjects in basic education. By adding digital security as a separate area, it would become visible and concrete in all areas of basic education.

This addition is well justified by the fact that digital security and its importance have grown in all sectors of society in recent years at a very fast pace. In addition, it is present everywhere in the school world. This is also evident in the open answers to the survey: most responses could not be determined to relate to a single subject. Rather, cybersecurity is visible in all subjects.

Some way of incorporating digital safety education into each subject would require additional resources for primary schools. In particular, adding an entire area to the concept of transversal competence would require extensive changes to the curriculum and, as a result, to the content of the subjects. To this end, teachers in both primary and lower secondary schools should be guaranteed access to continuing education in order to enable the inclusion of this component in their teaching.

Model 2: Digital security as part of the ICT competence area

A smaller structural change to make digital security more visible in all aspects of primary education would be to include digital security in the current ICT competence area.

Model 3: Digital security as part of extended ICT education

This objective could be achieved by strengthening the compulsory nature of ICT in primary and lower secondary schools in Finland. Currently, the availability of this elective subject depends on the school's own emphasis or willingness to offer ICT studies as elective courses. ICT education would include a digital security component.

Most of the teachers who responded to the survey stated that they do include cybersecurity in their teaching. If digital security were part of ICT education, it should be noted that this line of development also requires additional resources in terms of the content of teacher training, but also in terms of the number of staff.

4. General upper secondary education

4.1 Fundamentals of the general upper secondary education

Students who have successfully completed compulsory education are eligible for general or vocational upper secondary education and training. Student selection to upper secondary schools is mainly based on students' grades in their basic education certificate. More than 90 per cent of the relevant age group starts general or vocational upper secondary studies immediately after basic education. Completion of upper secondary education gives students eligibility to continue to higher education.

General upper secondary education provides, as its name suggests, general education. It does not qualify students for any particular occupation. At the end of general upper secondary school, students take a national school-leaving examination known as the Finnish matriculation examination. Those who pass the examination are eligible to apply for further studies at universities, universities of applied sciences and vocational institutions. General upper secondary education usually takes three years to complete. (FNAE, 2019)

The syllabus of general upper secondary education is designed to last three years, but students may complete it in 2 to 4 years. Instruction is organized in modular form not tied to year classes and students can decide on their individual study schedules rather freely. Each course is assessed on completion and when a student has completed the required number of courses, which include compulsory and elective studies, he or she receives a general upper secondary school certificate. (FNAE, 2019)

The Finnish National Agency for Education (2019) decides on the objectives and learning outcomes of the different subjects and study modules for general upper secondary education. Based on the national core curriculum, each education provider then prepares the local curriculum. Due to the modular structure of upper secondary education, students may combine studies from both upper general education and vocational education and training.

The curriculum was introduced locally as of 1 August 2021. While the new Act on General Upper Secondary Education entered into force on 1 August 2019, all aspects described in greater detail in local curricula, or those related to teaching, support for learning, guidance and cooperation, became obliging to education providers as of August 2021 (FNAE, 2019, 9).

The national core curriculum for general upper secondary education states that "Education and other activities in general upper secondary schools must be organized in accordance with the general national objectives defined in the Government Decree on General Upper Secondary Education (810/2018)." The aim is that the students may grow into educated members of society, acquire knowledge and skills required by the changing operating environment, and improve their capabilities for continuous learning. Particular emphasis is placed on the importance of transversal general knowledge and ability and understanding broad issues, and on encouraging the students towards ethically responsible and active agency as part of the local, national, European, and global community. (FNAE, 2019, 58.)

The students become accustomed to assessing the reliability of texts and information. In addition, the instruction guides the students in advancing their knowledge of information and communication technology and using it appropriately, responsibly, and safely, both when working alone and with others. (FNAE, 2019, 58.)

The areas of transversal competences comprise the common objectives of the general upper secondary school subjects (FNAE, 2019, 60). The transversal competences are complemented and expressed in concrete terms in the local curriculum for each subject and in the description of each study unit. Transversal competences are

taken into account in the school culture. Their implementation is complemented by descriptions of arrangements for familiarization with higher education studies and the world of work as well as international competence included in the curriculum. The contents of the thematic studies can be selected from the areas of transversal competences. (FNAE, 2019, 61–62.) Areas of transversal competences are:

1. Global and cultural competence
2. Well-being competence
3. Interaction competence
4. Multidisciplinary and creative competence
5. Societal competence
6. Ethical and environmental competence

4.2 Analysis of the research

The survey was targeted at 103 general upper secondary schools and sent to their principals. A total of 54 responses were received from 16 towns. The aim was to recruit respondents among teachers who include cybersecurity/information security/digital security in their teaching. Therefore, more than one teacher may have responded from each participating school.

Inclusion of cybersecurity in education in a timely manner as permitted by the subject

Teachers were asked to assess on a scale of 1 to 5 (completely disagree to fully agree) whether they include cybersecurity education as allowed by the subject requirements. The average score was 2.77. (See Fig 6)

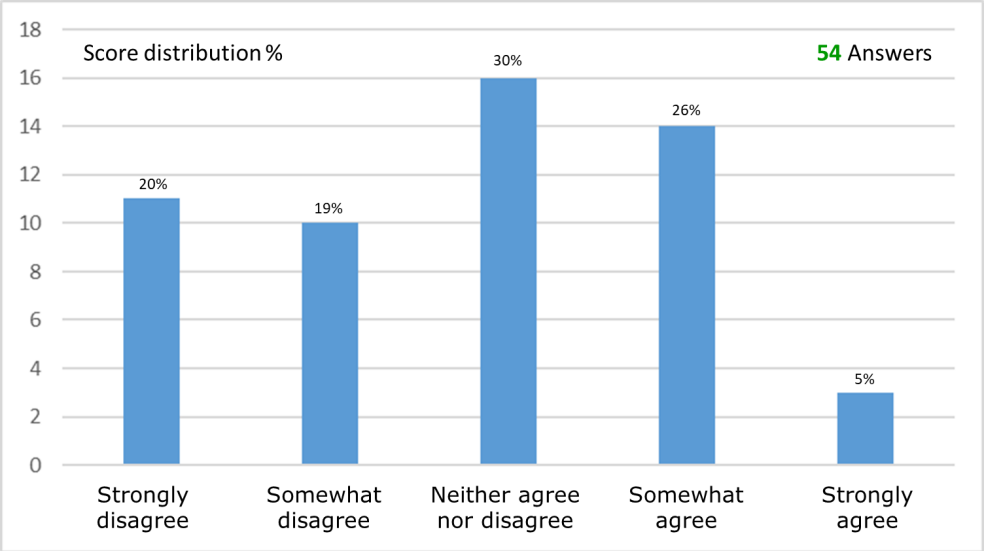


Figure 6: Inclusion of cybersecurity education in subject teaching.

Cybersecurity education should be included in every subject

The teachers were asked whether cybersecurity education should in some way be included in all subjects or whether it should be taught separately, for example, only as part of IT education. This divided teachers' opinions, also among ICT teachers. Taken together, the majority of all respondents (33) thought that cybersecurity belongs to all areas of the general upper secondary school. 21 respondents think it should be part of ICT teaching.

Clarity of the current curriculum guidelines for teaching cyber/information security

Teachers were asked to assess on a scale of 1 to 5 how clear the instructions in the current curriculum are in support of cybersecurity and information security education. The average was 1.96 (see Figure 7). The low average indicates that the responding teachers do not think the curriculum guidelines are clear when it comes to teaching cybersecurity and information security.

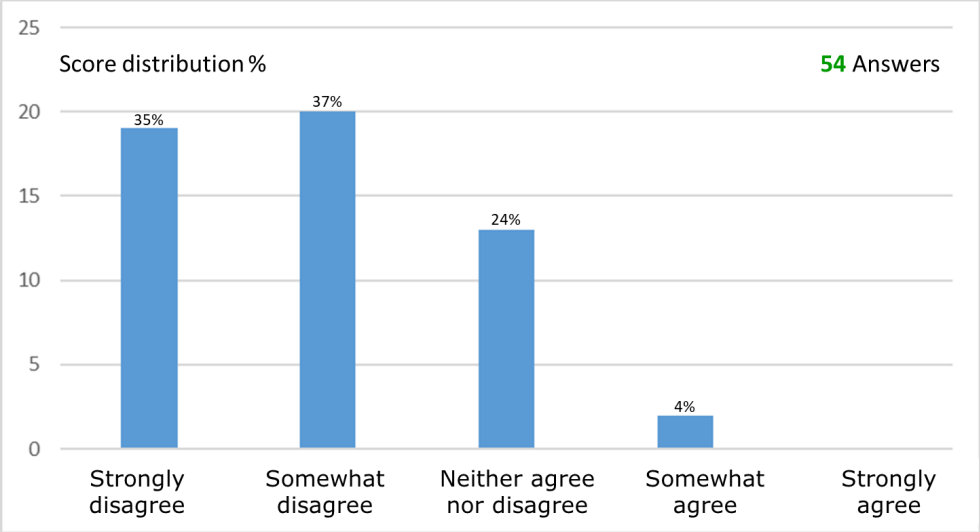


Figure 7: The clarity of the curriculum guidelines regarding cybersecurity teaching.

Teachers' capabilities in incorporating cybersecurity into teaching

On average, teachers were fairly neutral about their own ability to incorporate cybersecurity into their teaching. The assessment was carried out on a scale of 1 to 5. The average was 3.06.

4.3 Development of cybersecurity education in general upper secondary education

The development of cybersecurity education in general upper secondary education can be approached in several ways. This report explored teachers' views on whether cybersecurity education should in some way be included in each subject or whether it should be separated as part of a single subject such as information technology education. The study identified three models that are not mutually exclusive.

Model 1: Digital security for all areas of the general upper secondary school

This line of development could be achieved on the basis of the current curriculum by modifying the concept of transversal competences. Currently, transversal competence consists of six areas, and it forms the common objectives of all subjects. Adding digital security as an area of its own would make cybersecurity visible in all areas of the general upper secondary school. This addition is well justified by the fact that digital security and its importance have grown in all sectors of society. In addition, it is a multidisciplinary field of study, and this is also reflected in the survey responses.

Model 2: Digital security as part of the current areas of transversal competence

A smaller structural change to make digital security more visible in all aspects of general upper secondary education would be to include digital security in any of the existing competence areas.

Model 3: Digital security as part of ICT education

This line of development could be achieved by increasing the amount of ICT education in general upper secondary schools. Digital security would be included as one component of ICT education. The majority of responding ICT teachers stated that they already include digital security as part of their teaching. This line of development requires additional resources.

In addition, one line of development is to explore the possibility of creating general upper secondary schools specialised in the field of ICT in Finland, or alternatively the possibility of adding a specialisation into ICT to existing general upper secondary schools.

Incorporating digital safety education into all subjects would require additional resources for general upper secondary education. Especially adding an entire new area to the concept of transversal competence would require significant changes to the curriculum. Teachers should also be guaranteed access to continuous training if they so wish, so that the implementation of the component in the subject is equally possible for everyone.

5. Conclusion and discussion

The training needs of digital security and its importance are widely recognized. There is thus a willingness to develop training. Recent projects, such as the Cyber Security Development Programme, New Literacy Development Programme, and the Finnish National Board of Education's guidelines for schools on information security, demonstrate a desire to make digital security an important area in the planning of education and teaching at the level of basic education (FNAE, 2021). Currently, however, materials and tools serve more as guidelines and support than as obligations, so that responsibility for the use of these materials, or lack thereof, rests with the organizer of the teaching. In addition to ensuring that primary and lower secondary school pupils have digital security skills, there is also a need to ensure an adequate supply and level of education for teachers. This will cover the needs of the entire comprehensive school regarding digital security.

The national curriculum for general upper secondary education defines transversal competences and its different areas. These areas comprise the common objectives of the general upper secondary school subjects. Transversal competences create the preconditions for acquiring the knowledge and skills which enable students to cope with change in an increasingly digital and complex world. General upper secondary school education has several overall objectives defined in the curriculum. One of the objectives is to guide the individual in advancing their knowledge of information and communication technology and using it appropriately, responsibly, and safely, when working both alone and with others.

The results show that responding teachers do not feel that the curriculum contains clear instructions on how to teach cybersecurity. On average, teachers were fairly neutral about their own ability to incorporate cybersecurity into their teaching. Teachers' views were divided on whether cybersecurity education should in some way be included in all areas of the general upper secondary school, or whether cybersecurity education should be differentiated into a single subject such as information technology education. The majority of responding ICT teachers include cybersecurity education in their teaching.

Increasing and developing cybersecurity education can be approached from several different perspectives. This report identified different approaches which are not mutually exclusive. In the first approach, cybersecurity education will be integrated into all subjects by modifying the concept and content of transversal competences. The second option increases and develops cybersecurity education by improving access to ICT education and takes cybersecurity into account as part of its education.

IT skills, communication skills, media literacy and the use of social media constitute the foundation for the skills needed to use digital services. The proposed goals included the incorporation of ICT use as an integral part of the education in schools as well as in basic and supplementary teacher training. In addition, the investment in applied ICT know-how should be scaled up – with cyber security being one of its elements – and given a more prominent place in curriculum design throughout all education.

The challenge regarding the way things are now is the lack of cyber security education objectives for the entire education system. The sought-after improvement of competencies requires defining the basic skills and competencies for the entire national education system. It is necessary to have an understanding of what each citizen needs to know about cyber security, the demands of working life, what kind of professional skills are needed.

References

FNAE. (2014). Perusopetuksen opetussuunnitelman perusteet. https://www.oph.fi/fi/koulutus-ja-tutkinnot/tietoturva-ja-suoja-koulussa.

FNAE. (2019). National Core Curriculum for General Upper Secondary Education 2019, Finnish National Agency for Education. https://www.oph.fi/sites/default/files/documents/lukion_opetussuunnitelman_perusteet_2019.pdf.

FNAE. (2021). Tietoturva ja -suoja koulussa, Finnish National Agency for Education, https://www.oph.fi/fi/koulutus-ja-tutkinnot/tietoturva-ja-suoja-koulussa.

FNAE. (2022). Finnish education system, Finnish National Agency for Education. https://www.oph.fi/en/education-system.

Lehto M. (Edit.) (2022). Development Needs in Cybersecurity Education, University of Jyväskylä Research report 93/2022.

MEC. (2022). General education. https://okm.fi/en/general-education.

MTC. 2021. Finland's Cyber Security Development Programme 2021, Ministry of Transport and Communications, 2021:7.

Security Committee. (2019). Finland's cyber security strategy 2019, Government Resolution 3.10.2019.

VN. (2021). Education Policy Report of the Finnish Government, Publications of the Finnish Government 2021:64.

Russian Influence Operations during the Invasion of Ukraine

Joseph Littell, Nicolas Starck
Army Cyber Institute, West Point NY, United States

Joseph.Littell@WestPoint.edu
Nicolas.Starck@WestPoint.edu

Abstract: Prior to their invasion of Ukraine, the Russian Federation was seen as having a vastly superior ability to conduct operations in the information environment, particularly their ability to influence foreign audiences, when compared to their western counterparts (Cunningham, 2020). However, the efforts of Ukraine, NATO allies, and other aligned nations to conduct operations in the information environment with intelligence pre-bunking, traditional diplomacy, and sanctions, blunted many of Russia's best efforts. Further tactical and operational failures from the Armed Forces of the Russian Federation also undercut the salience of Russian messaging campaigns. This does not mean that Russia's efforts in the informational environment fully failed or did not adapt to these actions in the lead-up and continuation of the war. Through the use of the Natural Language Processing (NLP) technique transformer-based topic modeling (Grootendorst, 2022) and the causal inference technique Bayesian Structural Time Series analysis (Brodersen, 2015) this paper looks to both qualitatively and quantitatively examine how Russian state media on Twitter reacted, changed narratives, and focused efforts regionally from January 1st through September 1st, 2022. Through this analysis we argue that Russian efforts in Europe may have been of limited success. We further argue that by shifting focus, Russia gained influence in South America, and Middle East and North Africa, where their influence operations faced minimal obstructions, such as sanctions, and a latent anti-western sentiment.

Keywords: Influence Operations, Russia, Ukraine, Machine Learning, Natural Language Processing, Disinformation

Research Questions

1. How was Russia's state media messaging on Twitter oriented, both in frequency and content, prior to their invasion of Ukraine?
2. Was Russian messaging uniform, both in frequency and content, across languages and regions, or was it tailored by either?
3. Did Russian messaging change, either in frequency or content, in response to US and EU sanctions?
4. Did Russian messaging change, either in frequency or content, in response to changes in the physical/kinetic fight with Ukraine?

1. Background

The Russian Federation, like the Soviet Union preceding it, has used information in various ways to gain a strategic advantage both within its borders and with populations in foreign nations. Typically, this information is used to influence populations towards favorable positions for Russia. For example, during the First Chechen War Boris Yeltsin regularly lied to the Russian people. Starting as early as the day after Christmas, 1994, Yeltsin claimed that the military operations had ceased and Russia had successfully suppressed the Chechen rebellion (Hiatt, 1995). The historical record, however, proved that a further 18 months and tens of thousands of lives were lost before reaching that endpoint.

With improvements in telecommunications and the proliferation of the internet and social media, influence campaigns became much further reaching, with lower interaction costs and less impedance. Given their historical use of information, Russia was naturally one of the first nations to fully embrace the use of online influence campaigns as a mechanism to gain support for their military endeavors. This first materialized in a noticeable way during the annexation of Southern Ossetia from Georgia in 2008. Russia used a measured, now familiar, narrative of only acting defensively to protect ethnic Russians from Georgian aggression which was distributed through Russian outlets, to include web-based news and social media (Allison, 2008). The decentralized and pervasive nature of the campaign generated the sufficient ambiguity and plausible deniability that national governments and the international community turned a relatively blind eye to the incursion. The phenomenon of online information operations accelerated in 2014, again acting as a force multiplier to create the freedom of maneuver for Russia to conduct its invasion and annexation of Crimea.

Following the revelations that Russian influence campaigns targeted the 2016 US Presidential Election (National Intelligence Council, 2021) and the 2016 European Union Referendum in the United Kingdom (Bastos and Mercea, 2018), Russian online influence campaigns were seen as the pinnacle of information warfare. The disclosures relating to the 2016 US election led to several investigations which ultimately confirmed Russian

involvement. The investigations showed that Russia had directly targeted civilian populations, including influencing the conduct of local rallies by US citizens in support of their preferred political candidates. Regardless of the direct electoral efficacy of these influence operations on the US and UK populations, the operations created information effects. Numerous media outlets and think tanks opined that Russia was far outpacing their western counterparts' ability to conduct strategic information warfare, principally online influence campaigns. However, following the 2022 invasion of Ukraine much of the mystique and rhetoric surrounding Russian capability has faltered. Deliberate pre-bunking efforts and a unified NATO helped blunt Russia's efforts to sow doubt and dissuade Western support for the defense of Ukraine. In this paper, we present an analysis of the impacts of some of these counter-efforts and their effect on Russian information operations.

2. Methodology

The tweets used for this study were gathered by the Alliance for Securing Democracy for their Hamilton 2.0 Dashboard using the Twitter API (Schafer, 2019). A total of 611,860 tweets were compiled from 344 screen names associated with the Russian government. This set of screen names included accounts of news media sources such as Sputnik, RT, Tass, their various regional publications, and individual reporters associated with those outlets. The data sample includes 53 languages collected from January 1st, 2022 until September 1st, 2022. All tweets were machine translated to English from their native language through IBM Watson's translation.

We then filtered the data down to seven specifically selected languages: Russian, English, Spanish, French, Arabic, Turkish, and German. We chose these languages for four reasons. First, Russian, English, Arabic, Spanish, and Turkish, in order, are the five most prevalent languages within the data set. German and French round out the top ten languages, at nine and ten, respectively. Second, Arabic, Spanish, and to a lesser extent French, can be used to reflect Russian regional messaging in the Middle East, Central and South America, and Francophone Africa. Third, English, French, German, Spanish, and Turkish are all languages of NATO nations, a critical audience for the conflict. Finally, the Russian language allows for the examination of the messaging efforts about the invasion of Ukraine within the domestic population and the Russian diaspora. This filtered data set included 532,510 tweets from 341 accounts.

We used two approaches to answer the stated research questions. First to use quantitative methods to guide a qualitative assessment of the content, and second to quantify the frequency of the messaging. Both approaches used established machine learning techniques (Grootendorst, 2022) (Brodersen, 2015). As such, the methodology will summarize those methods and describe their application to answer our research questions.

We first conducted an analysis of the content of the filtered collection of tweets using Maarten Grootendorst's BERTopic: Neural topic modeling with a class-based TF-IDF procedure (Grootendorst, 2022). This methodology uses transformer based neural networks to convert sentences into a vector space that can be used to compare sentences semantically. Sentences are converted into an array of numbers that represent their position in some multi-dimensional space. This allows for the relationships between sentences to be measured by their distance to other distinct sentences within that same multi-dimensional space. Similar sentences will be closer in distance to each other, reflecting similar content. Further description of this process can be found in Devlin et al (2019) seminal paper introducing the Bidirectional Encoder Representations from Transformers (BERT) model. Comparing these large, multi-dimensional vectors can be computationally and temporally expensive. This can be mitigated by several dimensionality reduction tools within the field of machine learning. We selected Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) for its ability to retain global features projected onto a lower dimension. Once vectors are reduced, they are clustered based on the density in which they appear in the reduced space through a Hierarchical Based Density Scan (HBDSCAN) (Campello, 2013). This process allows for the identification of the most significant clusters within vector space. Finally, each density-based cluster is expressed as a human-understandable topic using Term Frequency Inverse Document Frequency (TF-IDF) to assign the most distinct words with a cluster as its identifier.

Our second approach applied a causal inference analysis to the frequency of Russian messaging on Twitter using Bayesian structural time-series models (Brodersen, 2015). We counted tweets by day to visualize the frequency of messaging over time. This daily count was measured from January 1st through September 1st, 2022. We then processed these daily counts through a diffusion-regression state-space model to calculate the expected quantity absent an intervention. This baseline projection allows for the inference of the temporal changes that are explained by the impact of a given intervention, the incorporation of Bayesian priors during the treatment, and to account for many sources of potential variation including the time-varying covariates. We can then

calculate the difference between the expected volume of Russian messaging given no intervention and the actual value to determine if there is statistically significant difference that can be correlated with the intervention. This metric provides a quantitative measure of the impact of these intervention strategies on Russian messaging. We then identified a selection of dates associated with interventions, such as waves of sanctions, changes to military strategy, and debt default, for which we could quantitatively measure the potential impact to Russian messaging. These events were selected based on the hypothesis that they would change the frequency of Russian activity on Twitter. While proving true causal relationships is not possible with the available data, this approach allowed us to quantify the effect of particular counter-influence actions relating to the Russian invasion of Ukraine.

3. Results

3.1 RQ1. How was Russia's state media messaging on Twitter oriented, both in frequency and content, prior to their invasion of Ukraine?

Prior to Russian Federation Forces crossing into Western Ukraine across multiple avenues of advance, Russian outlets and political leaders attempted to downplay the belief that Russia was amassing forces for a full-scale invasion. Russian Forces were being moved to the Ukrainian border throughout 2021, to such an alarming degree that on March 31st, 2021, the United States European Command raised their awareness level to indicate a potential crisis (Funaiole et al, 2021). By December 2021, Russia demanded a reduction of NATO presence in Central and Eastern Europe and a permanent restriction against Ukraine ever joining NATO (Wintour, 2022). These actions occurred immediately the start of our data sample on January 1st, 2022.

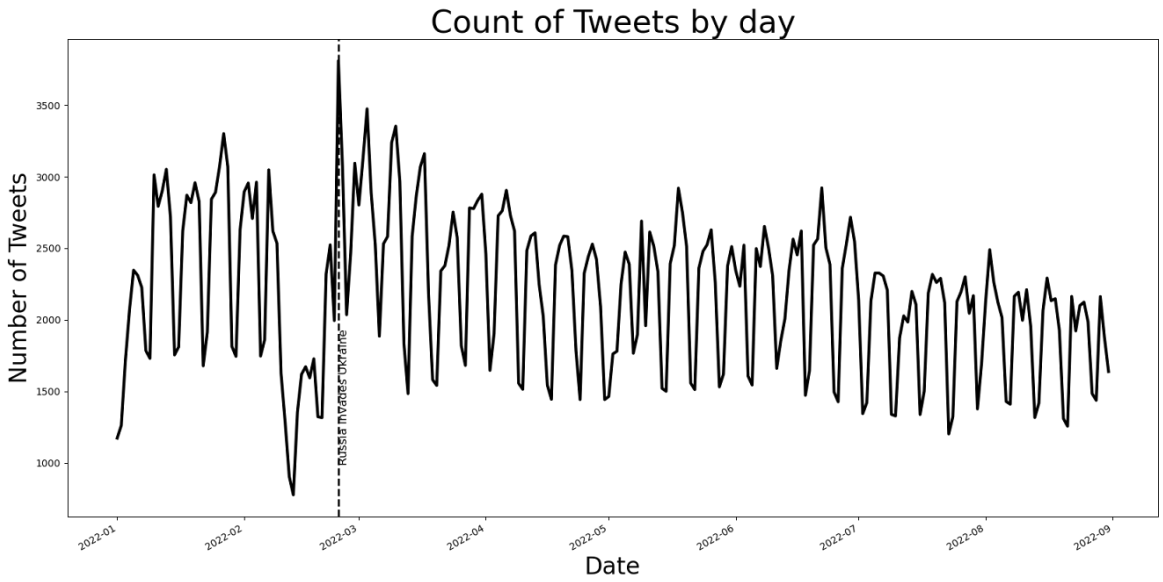


Figure 1: Frequency of Russian Media Tweets from January 1st through September 1st, 2022

The frequency of posts among all languages was stable until the February 4th, 2022, when all languages saw a sharp decline. The lowest volume of Russian Twitter activity was reached on February 13th, 2022, with only 778 tweets across all languages as compared to a daily average of 2,191 tweets. This decline occurred simultaneously with a series of key pronouncements from the West: (1) the deployment of 3,000 more NATO troops within Romania, Poland, and Germany, (2) coordinated public announcements of unified efforts to deter Russian aggression, and (3) the intent to impose severe economic sanctions if Russia intervened militarily in Ukraine. This decline, however, was not uniform across all languages. German language tweets increased during this period, and the decline in English language tweets was smaller than the other languages. This low point in Russian Twitter activity was followed by a sharp increase across all languages leading up to the invasion, with the highest volume, 3,810 tweets, occurring on February 24th. This represented a 74% increase on the average number of tweets across the dataset.

As with the frequency, there was a drastic shift in narrative and language in this same period preceding the invasion. Prior to the invasion there was no discernible consistent Russian narrative in any of the languages in

our sample. This pattern shifted around the launch of the "special military operation". We now present a qualitative summary of NLP analysis by region, however the higher order trend that appeared was Russia trying to hide their intentions to invade Ukraine until approximately February 13th, at which point numerous regionally aligned narratives surfaced. These emergent narratives served one goal - to justify Russian military actions within Ukraine.

3.1.1 Russian

Within the Russian language tweets in the dataset preceding the invasion, we saw a range of regular news topics. This spanned domestic news such as COVID-19, the success of Russian-born athletes in the 2022 Winter Olympic Games, and international reporting on events like the Kazakhstan protests in early January and the Tonga Volcano and subsequent Tsunami. Notably absent was news or narratives relating to the massive troop buildup along the Ukrainian border.

However, in the surge of activity in the lead up to the invasion, Ukraine appeared as the central topic framed in two ways. The first framed the action as a large deterrence exercise aimed at showing the West Russia's military might. The second framed a quick military intervention against Ukraine as critically needed due to their alignment with Nazis, and the needed recognition of the eastern Ukrainian regions of Lugansk and Donetsk as independent states.

3.1.2 Americas

Within the English-speaking world, the narrative predominantly focused on various COVID-19 topics including the Sputnik vaccine and its superiority to the western mRNA vaccines, Western vaccine mandates, and other anti-vaccine conspiracies. These COVID-19 narratives were coupled with news coverage of Julian Assange's extradition, flooding in Brazil, and sporting events like the UFC. Mentions of Ukraine were sparse, only spiking once following US Secretary of State Anthony Blinken's reaffirmation of Ukraine's sovereignty being essential to global order (Zhang, 2022).

Following the decline in mid-February relative to the previous 6 weeks, the narrative changed drastically in the lead up to the invasion. Multiple major conspiracy theories appeared with the goal to add legitimacy to the forthcoming invasion or at worst, erode western support to Ukraine. The first stipulated that Dominion Voting Systems servers – allegedly used to rig the 2020 US Presidential election in favor of now-President Joseph R. Biden – were being held in Ukraine. The second conspiracy was that bioweapons, to include SARS-CoV-2, were engineered in US-backed labs across Ukraine. Another major conspiracy was that Ukrainian President Vladamire Zelenskyy was a puppet of the West, particularly the United States, and therefore an illegitimate leader. Finally, the persistent narrative that Ukraine has always been a part of Russia, the so-called country and culture is not real, and therefore Russia has historical rights to the land.

In Central and South America, the Spanish language tweets were again tailored to regional news. COVID-19 once again was a major source of content, as was traditional news including the revolt in Kazakhstan, an oil spill in Peru, and ongoing political strife and violence in Honduras. Like the English language tweets, US Secretary of State Anthony Blinken's affirmation of Ukrainian sovereignty was refuted by the Russian media. However, the Spanish language messaging directly labeled the statement as disinformation and included a fact sheet to counter Secretary Blinken's statement. The other mentions of Ukraine framed the Russian activity in several ways. First as a simple military exercise which should not warrant a NATO response, and second as necessary to evacuate Donetsk and Lugansk due to Ukrainian aggression. This pre-invasion rhetoric culminated with the announcement of a Special Military Operation to ensure the independence of the newly formed breakaway republics.

3.1.3 Middle East and North Africa

Arabic themes in January and early February centered around news and local events of the Middle East and North Africa (MENA). Topics included the ongoing conflict between Saudi Arabia and Houthi rebels in Yemen, the African Cup soccer tournament and Burkina Faso's championship victory, ISIS reconstituting in Syria, and a looming Libyan government collapse. News of COVID-19, prevalent in many other languages' tweets, were minimal within the Arabic language data. The Arabic themes mirrored the pattern seen in the Russian language tweets. Ukraine content was nearly non-existent prior to the decline in overall volume, followed by the narratives that Russia was a benign actor resisting Western aggression and the necessity for Russia's recognition of Lugansk and Donetsk as independent states.

As Turkey sits at the crossroads of Asia and Europe, its news content was a mix of the MENA content and the COVID-19 narratives in the western language data. Another common narrative focused on Turkey's dire

economic situation with rising unemployment, runaway inflation, and rising energy costs. No mention of Ukraine existed before the lead-up to the Russian Invasion, to include mentions of US Secretary of State Blinken's statements in January 2022. Curiously, mentions of Ukraine were still minimal in the lead up to the invasion. However, the messaging that was present followed the previously established narrative of protecting the newly independent Donbass republics from Ukrainian aggression.

3.1.4 Europe

French language tweets focused on traditional news, such as COVID-19 and travel-related content, with minimal mention of the Ukrainian or the troop build-up. The messaging centered on the then-upcoming French Presidential election prior to the invasion, potentially in the hopes of casting support for Ukraine as politically toxic within the French population to dissuade further support by political leaders. This lack of content relating to Ukraine remained consistent throughout the ramp up to the invasion.

The German language tweets ran contrary to all the other observations. The German tweets saw no decline in the frequency of posts in the run up to the invasion. The main narrative structure focused on the rising costs of energy and Nord Stream 2 construction. As with the assumption that political motivation may have deterred French actions, the economic costs and Germany's dependence on imported petroleum seemed to be the focal point for Russian messaging targeted at German audiences.

3.2 RQ2. Was Russian messaging uniform, both in frequency and content, across languages and regions, or was it tailored by either?

As demonstrated in the analysis of our first research question, Russia tailored its messaging towards specific regions. While Russia content was consistent in seeking ambiguity with regards to its invasion of Ukraine, the motivations and messaging justifying the necessity of the military intervention differed greatly by region. Regions outside of the West had their stories biased toward Western military aggression and imperialism. This framing had the potential to resonate in regions where distrust of the West is high or regions accustomed to Western military interventions, such as the Spanish, Arabic and Turkish speaking world. Russian speakers, while still susceptible to similar narratives, may have required further reasoning why military intervention, rather than standard deterrence, was needed. This could explain the rhetorical focus on Ukraine's supposed Nazism, the perpetration of Ukrainian war crimes in the Donbass War, and the importance of intervening militarily to ensure the safety and self-determination of the Russian-speaking regions of Ukraine.

This contrasts starkly with the Western messaging, which had vastly different narratives and rationales concerning the purpose of the invasion. The English language tweets focused on vaccine hesitancy, conspiracy, and US-specific political motifs. The German language messaging focused on industry and the economy, particularly the energy sector. Finally, the French content focused on appeals to freedom and independence from tyrannical Ukrainian rule. While the Western tweets paint NATO and the EU in imperialistic terms, it does not belabor the point. This is an intuitive observation, as the West would be unlikely to view their own history of interventionism as a negative. Therefore, such messaging would be counterproductive in the pursuit of dissuading a response in the defense of Ukraine.

3.3 RQ3. Did Russian messaging change, either in frequency or content, in response to US and EU sanctions?

The day following the February 21st, 2022, national television address by Russian President Vladimir Putin, Western nations announced their first round of economic sanctions against Russian political leaders and oligarchs. By the 26th of February, these sanctions escalated from individuals to Russian institutions including the banking sector, access to western social media, internet services, oil imports and exports, and more. While additional sanctions continued to be imposed throughout the conflict, the vast majority were in place within the first month of the conflict. The most temporally direct effect to Russian information campaigns came on February 27th, 2022, when the European Union member states banned all Russian state-owned outlets, such as RT and Sputnik (Siebold and Blenkinsop, 2022).

Taking the first five weeks of data collection as a baseline, and accounting for the lull in tweets from February 8th to the 23rd, Russia media overall saw a steady decline since the invasion in its output of content across all languages combined as shown in Figure 2 below. However, not every language decreased at the same rate nor degree. For example, English, French, and German all saw a significant, steady decline, while Arabic tweets had

a minimal decline, and Turkish and Spanish language tweets both had periods of increasing frequency following the invasion and requisite sanctions.

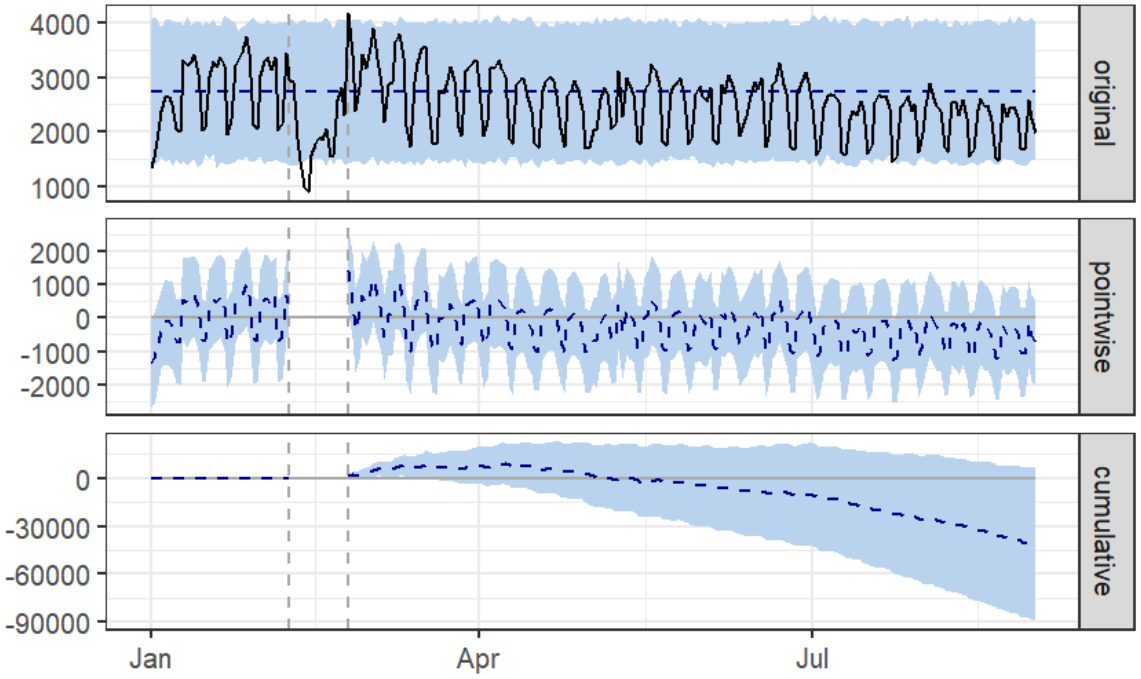


Figure 2: Impact of sanctions of Russian Media Tweets across all languages

During the post-intervention period, the average number of daily tweets was approximately 2,520. In the absence of any sanctions, we would have expected an average response of 2,740. In relative terms, the actual average number of daily tweets, our response variable, showed a decrease of 8% with the 95% interval of this percentage being [-17%, +1%]. This is a statistically significant finding as the likelihood of this happening by chance is very small, with the Bayesian one-sided tail-area probability p = 0.048. Comparing the cumulative effect against the French and German data, where these sanctions took place, the decline becomes even more clear. During the post-intervention period, French language tweets had a daily average value of approximately 34.20. In the absence of an intervention, we would have expected an average of 69.38 tweets per day. This is a 51% decline, with a 95% interval of this percentage is [-61%, -41%] and a Bayesian one-sided tail-area probability p = 0.001. German language tweets saw an even greater effect than French, with a 54% [-76%, -34%] decline with a Bayesian one-sided tail-area probability p = 0.001.

By contrast, for Spanish language tweets we would have expected an average response of 374.31 in the absence of an intervention. However, during the post-intervention period the response variable had an average daily value of approximately 352.91 daily tweets. This is only 6% lower than the prior expectation with a 95% confidence interval of [-17%, +5%]. Observing Figure 3 below, we can be seen that a drop off did not occur until July, potentially as a result of the official recognition on July 27th of Russia's debt default from the prior month (King and Jordan, 2022). Unlike the overall, French, and German examples, this finding is not statistically significant as it has a Bayesian one-sided tail-area probability p = 0.148. This is particularly notable compared to the Arabic dataset, another language unaffected by direct sanctions. The Russian activity on Arabic language Twitter saw only a 7% decline overall, with a 95% confidence interval of [-13%, -0%]. These findings, unlike those for Spanish, were statistically significant with a Bayesian one-sided tail-area probability p = 0.019.

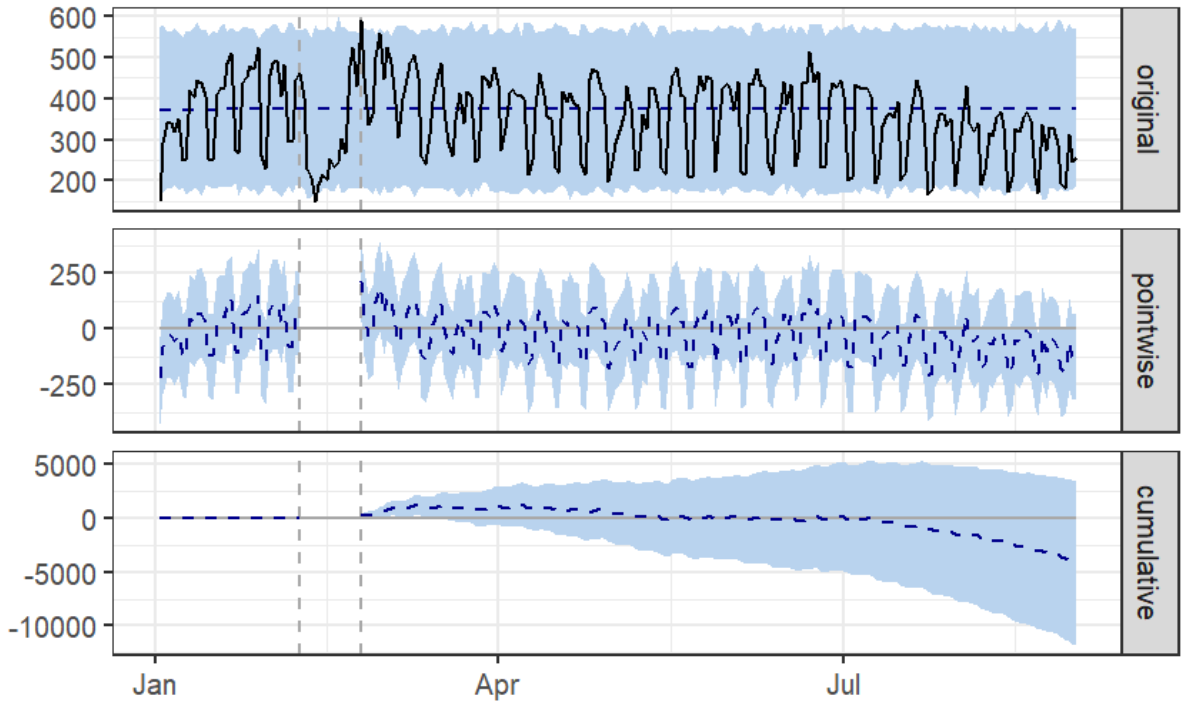


Figure 3: Impact of sanctions of Russian Media on Spanish language tweets

Following the EU sanctions of Russian state-run media, two narratives appeared in the remaining French and German language tweets. The first narrative was a complaint about censorship of Russian media. This narrative pointed out the hypocrisy of the West in removing Russian media outlets access to these platforms and made claims that Western governments control Western social media platforms . The second narrative tried to migrate users from Twitter onto Telegram. This migration would allow the uncensored expression of Russian narratives, and interaction within the Telegram channel could be monitored and moderated by Russian actors. By contrast, many of the same narratives used in the lead up to the invasion persisted well into June 2022 within the English language data. The narrative around biological labs was particularly active throughout the spring and summer before being surpassed by a narrative blaming Ukrainian support for rising inflation.

3.4 RQ4. Did Russian messaging change, either in frequency or content, in response to changes in the physical/kinetic fight with Ukraine?

From the standpoint of frequency, we found no statistically significance date from a set of the following events when looking at the cumulative data set – the Bucha massacre (2022-04-01), the end of Kiev Offensive (2022-04-04), the launching of Eastern Offensive (2022-04-18), Finland and Sweden signal interest in joining NATO (2022-04-10), Finland and Sweden officially apply for NATO (2022-05-18), the Russian capture of Mariupol (2022-05-20), Russia failing to pay foreign debt (2022-06-24), or Russia officially declared to have defaulted on its debt (2022-06-27).

However, when separated by language, some trends do appear. As mentioned in RQ3, Spanish language saw a decline in posts following the Russian foreign debt default, which began on June 24th, 2022. This was a relative decline, visualized in Figure 4 below, of 16% with a 95% confidence interval of [-24%, -8%], with a Bayesian one-sided tail-area probability p = 0.001. Russian language activity saw a similar decline of 6% [-19%, +6%] although it was not statistically significant (p = 0.176). No other event causes a noticeable change nor one that would be statistically significant.

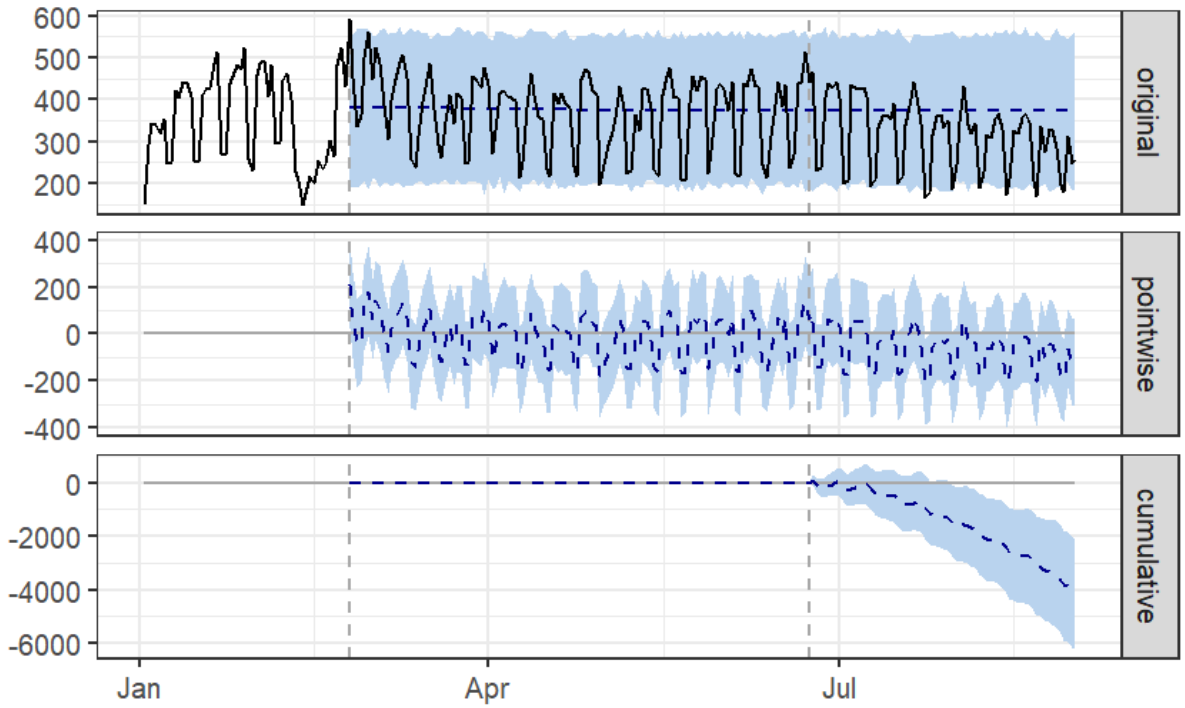


Figure 4: Impact of Debt Default on Spanish language tweets

For content, Russia references all of these events and used them as part of regionally based narratives. Within the English dataset, the messaging responded to the Bucha massacre claiming it as a hoax with actors and staged images. Within the Russian language data, the massacre was also labeled as a hoax. Notably this narrative also claimed if the event were not a hoax, a Nazi presence in Bucha were the real perpetrators. The messaging response to the capture of Mariupol also reflected subtle differences between languages. In Russian, the narrative focused on solidifying support for the "Special Military Operation," fighting off the Nazis within the Azoz Battalion, and ties to the WW2 victory against Nazi Germany. In English, however, the narrative focused on Ukrainians committing war crimes against the civilian population.

4. Conclusion

Numerous factors affected the Russian information campaign surrounding the invasion of Ukraine and its later narrative alignment towards distinct regional audiences. Previously, Russia had a great deal of informational freedom and were relatively uncontested when dealing with the West. In certain cases, such as the 2016 US Presidential Election and the 2016 European Union Referendum in the United Kingdom, Russia tailored its narratives to issues with deep seeded cultural significance, allowing it to compound existing social and cultural rifts. Russia's information operations supporting military actions in Georgia and the 2014 annexation of Crimea existed in a vacuum, with limited push back from opposing informational forces. Any Western push back that did occur was limited and often not coordinated among allies, leaving large gaps for Russian narratives to permeate. Notably, the speed with which these military actions were concluded also limited the time for a response to develop.

However, Russia's invasion of Ukraine faced a unified front from the West. This primarily occurred in the form of pre-bunking Russian narratives and levying heavy sanctions which hindered some informational capabilities as shown in RQ3. Many traditional military functions including logistics were also degraded. This forced Russia to rely on long shuttered factories, mothballed equipment, outside industry, and coordination with the few allies it had remaining. Some of the sanctions directly hindered capabilities by attacking Russian influence at its source within a given country, such as European Union member nations shutting down RT and Sputnik (King and Jordan, 2022). Other sanctions degraded non-informational aspects of the invasion, which we speculate may have caused Russia to shift support or funding away from informational efforts.

These sanctions packages, along with Russia's own failures on the battlefield, slowed the pace of the war and eroded any previous first mover advantage Russia may have had. These delays also allowed for Ukraine to deploy a strong information capability to counter Russian narratives. Ukraine's information campaign was bolstered by examples of both military victories and Russian war crimes. The long buildup of Russian forces and length of the

war also allowed traditional new organizations across the globe to send journalists and war correspondents to Ukraine. This effort added more visibility and first-hand content from the war that were not present during the quick invasions of Georgia in 2008 and Crimea in 2014. Taken as a whole, the global response to the Russian invasion of Ukraine in 2022 demonstrated an increased understanding and sophistication of the information environment. This should prompt a re-evaluation of the established view that Russia is the preeminent force in the information environment and serve as a roadmap for future effective responses to their tactics.

References

Allison, R. (2008). Russia resurgent? Moscow's campaign to 'coerce Georgia to peace.' The Royal Institute of International Affairs. Retrieved October 12, 2022, from https://commonweb.unifr.ch/artsdean/pub/gestens/f/as/files/4760/39349_201918.pdf

Bastos, M., & Mercea, D. (2018). The public accountability of social platforms: lessons from a study on bots and trolls in the Brexit campaign. The Royal Society Publishing. Retrieved October 12, 2022, from https://doi.org/10.1098/rsta.2018.0003

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015, January 1). Inferring causal impact using Bayesian structural time-series models. Google Research. Retrieved October 12, 2022, from https://research.google/pubs/pub41854/

Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Cunningham, C. (2020, November 18). A russian federation information warfare primer. The Henry M. Jackson School of International Studies. Retrieved October 12, 2022, from https://jsis.washington.edu/news/a-russian-federation-information-warfare-primer/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. arXiv.org. Retrieved October 12, 2022, from https://arxiv.org/abs/1810.04805

Funaiole, M., Bermudez J., Conley H., & Newlin C. (2021, April 22). Unpacking the Russian troop buildup along Ukraine's border. Center for Strategic and International Studies. Retrieve October 12,2022, from https://www.csis.org/analysis/unpacking-russian-troop-buildup-along-ukraines-border

Grootendorst, M. (2022, March 11). Bertopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv.org. Retrieved October 12, 2022, from https://arxiv.org/abs/2203.05794

Hiatt, F. (1995, January 29). Truth of chechen war derails Russian Propaganda Machine. The Washington Post. Retrieved October 12, 2022, from https://www.washingtonpost.com/archive/politics/1995/01/29/truth-of-chechen-war-derails-russian-propaganda-machine/62c813f3-9f8d-47e1-bed1-254443ab3d69/

King, B., & Jordan, D. (2022, June 27). Russia in debt default as payment deadline passes. BBC News. Retrieved October 12, 2022, from https://www.bbc.com/news/business-61929926

National Intelligence Council (2021). Foreign Threats to the 2020 US Federal Election. Retrieved October 12, 2022, from https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf

Schafer, B. (2019, September 3). Hamilton 2.0 methodology & faqs. Alliance For Securing Democracy. Retrieved October 12, 2022, from https://securingdemocracy.gmfus.org/hamilton-2-0-methodology-faqs/

Siebold, S., & Blenkinsop, P. (2022, February 27). EU tightens Russian sanctions and buys weapons for Ukraine. Reuters. Retrieved October 12, 2022, from https://www.reuters.com/world/europe/eu-close-airspace-russia-curb-media-target-belarus-2022-02-27/

Wintour, P. (2022, January 20). Blinken to reassert Ukraine sovereignty in Berlin speech. The Guardian. Retrieved October 12, 2022, from https://www.theguardian.com/us-news/2022/jan/20/antony-blinken-us-ukraine-russia-berlin-speech

Zhang, Z. (2022, March 30). Study Confirms Influence of Russian Internet "Trolls" on 2016 Election. Columbia SIPA. Retrieved October 12, 2022, from https://www.sipa.columbia.edu/news/study-confirms-influence-russian-internet-trolls-2016-election

Modelling the Influential Factors Embedded in the Proportionality Assessment in Military Operations

Clara Maathuis[1] and Sabarathinam Chockalingam[2]
[1]Open University of the Netherlands, Heerlen, The Netherlands.
[2]Institute for Energy Technology, Halden, Norway.
clara.maathuis@ou.nl
sabarathinam.chockalingam@ife.no

Abstract: The ongoing decade was believed to be a peaceful one. However, contemporary conflicts, and in particular, ongoing wars prove the opposite as they show the increase in context complexity when defining their goals as well as execution strategies for building means and methods for achieving them by gaining advantage against their adversaries through the engagement of well-established targets. At the core of the engagement decision relies the principle of proportionality which brings in a direct relation the expected unintended effects on civilian side with the anticipated intended effects on military side. While the clusters of effects involved in the proportionality assessment are clear, the process itself is subjective, governed by different dimensions of uncertainty, and represents the responsibility of military Commanders. Thus, a complex socio-technical process where different clusters of influential factors (e.g., military, technical, socio-ethical) play a role in the decisions made. Having said that, the objective of this research is to capture and cluster these factors, and further to model their influence in the proportionality decision-making process. This decision support system produces military targeting awareness to the agents involved in the processes of building, executing, and assessing military operations. To accomplish the aim of this research, a Design Science Research methodological approach is taken for capturing and modelling the influential factors as a socio-technical artefact in the form of a Bayesian Belief Network (BBN) model. The model proposed is further evaluated through demonstration on three different cases in respect to real military operations incidents and scenarios existing in the scientific literature in this research field. Hence, through this demonstration, it is illustrated and interpreted how the factors identified influence proportionality decisions when assessing target engagement as being proportional or disproportional. In these cases, corresponding measures for strengthening proportionality and reducing disproportionality in military operations are considered.

Keywords: Military Operations, Targeting, Proportionality, Cyber Operations, Machine Learning, Bayesian Networks.

1. Introduction

"When you change the way you look at things, the things you look at change." (Max Planck)

Recent and ongoing conflicts prove the fact that technological innovation and the integration of technological developments through advanced, adaptive, and/or intelligent means and methods of warfare, imply a higher degree of awareness and from there involvement when aiming to achieve well defined military objectives (the what) while building strategies and plans to accomplish them in respect with the necessary conditions and limitations (the how). No matter if conventional or unconventional instruments of war are being used and no matter in which operational domain (e.g., land, air, cyber), at the core of this phenomenon is the targeting process where the indispensable element in the military decision-making process when deciding if it is proper to engage a military target or not is the *proportionality assessment*. This assessment brings together two cardinally opposed components, i.e., military advantage and collateral damage (Oxford Institute for Ethics, Law, and Armed Conflict, 2009; Dinstein, 2016; McKenna, 2020).

While this assessment takes place considering the information available, it is a subjective process where the decision is made by the military Commander who has to refrain from attack in case that the incidental civilian harm is excessive in relation to the anticipated military desired effects (Preston & Taylor, 2016). In this process, a series of influential factors (of both military and human nature) play a role (Maathuis, Pieters & van den Berg, 2021). While a large body of knowledge focused on understanding the legitimacy of target engagement and the two major components involved, limited studies focused on gathering the influential factors that could play a role when conducting proportionality assessment, and such a fact could imply major and large-scale consequences on corresponding civilian and military systems and humans, and further negatively influence the mission.

Hence, the aim of this research is to identify, structure, and model various influential factors to proportionality decisions by means of a BBN (Bayesian Belief Network) model and further demonstrate the model proposed on three influencing situations. In this regard, to the best of our knowledge, this model represents the first attempt to capture and model the influential factors to proportionality decisions in military operations. Accordingly, the contributions of the model proposed are defined as follows:

- Awareness to policy and military decision makers such as military Commanders and military-legal experts on the existence of diverse influential factors to proportionality decisions.
- Support for the design and development of intelligent or software-based solutions that could assist military decision makers when conducting the proportionality assessment and further considering relevant aspects to other activities such as CoAs (Courses of Action) definition, analysis, and comparison.
- A call for further active research in this scientific domain by merging scientific discourses and means with practical perspectives and methods to assure the development, execution, and assessment of transparent and responsible military operations.

To achieve the above mentioned research goals and sustain these contributions, multidisciplinary research is carried out based on extensive literature review and model development following the Design Science Research methodology.

The reminder of this article is structured as follows. Section 2 presents the context of this research and discusses relevant studies. Section 3 explains the research methodology followed to achieve the aim of this research. Section 4 depicts and clusters the influential factors to proportionality decisions identified. Section 5 discusses the design of the model proposed in this research, i.e., variables and relations between these variables. Section 6 presents the evaluation of the model by means of demonstration of three different influence cases, and interprets the results obtained. At the end, Section 7 provides the findings of this research and addresses future research ideas.

2. Research Background and Related Research

When targeting in military operations, effects must be created to achieve the military objective(s) defined and reach the end state of the mission. This process is carried out under the guidance of the governing law, i.e., International Humanitarian Law (IHL) and the Rules of Engagement (RoEs) specifically defined for a military operation. Herein, military targets are selected, and further the anticipated military advantage and the expected collateral damage are established to decide if the attack is proportional or not, i.e., can be approved or not for execution (Henderson & Reece, 2018; Ali, 2020). This process is based on the principle of proportionality, is referred in the literature with terms such as assessment or test and has a subjective nature as it is based on human reasoning and responsibility of the military Commander in place (McKenna, 2020). Specifically, the principle of proportionality forbids the expected total incidental civilian loss (i.e., collateral damage) to be excessive in relation to the direct anticipated military advantage (AP I Art. 51(5)(b), 1977; NATO, 2016). Hence, in the proportionality assessment a balancing or comparison between antagonistic categories which reflect the 'interest of the belligerent in carrying out a military action and the interest of civilians, who, although extraneous in the conduct of hostilities, might be victimized by that action' (Cannizzaro, 2006), in other words, the intentional effects or the positive part of the (in)equation versus the unintentional effects or the negative part of the equation (Oxford Institute for Ethics, Law, and Armed Conflict (2009); Maathuis, Pieters & van den Berg, (2016); Maathuis, Pieters & van den Berg, (2018a)). As proportionality is not the singular process to be considered when preparing and executing a military operation (Gillard, 2018), so is the existence of other factors that belong to the human realm (e.g., background and experience of the military Commander). This also plays an influential role in the military decision-making processes involved (NATO, 2016; Henderson & Reece, 2018; Maathuis, 2022a).

Hence, we further consider the multidisciplinary nature of this research and discuss related studies that have investigated different types of influential factors or variables surrounding and playing a role when the proportionality assessment is carried out. On this behalf, Henderson & Reece, (2018) scrutinize what a 'reasonable military Commander' means in the context of proportionality assessment and consider that this assessment (i) is subjective when the person believes, (ii) is objective but unqualified when the person reasonably believes, and (iii) is objective but qualified when an expert is reasonable objective. To this end, a reasonable military Commander is a 'reasonably well-informed person in the circumstances of the actual perpetrator, making reasonable use of the information available' (Gall, 2004) which reflects the importance of military training and exercise that facilitate proper reasoning and explanations of the decisions taken (Ali, 2020; Maathuis, 2022b). Furthermore, Hayashi (2010) considers that proportionality decisions could be taken in very stressful circumstances, illustrates the importance of the military purpose that must be achieved, and points out the negative role that poor information could have for proportionality assessment. At the same time, Fenrick (2001) reflects on the role that the background and values of the decision-maker have when conducting

proportionality assessment and pose critical questions regarding inclusion and exclusion criteria plus temporal and spatial factors that should be included when conducting proportionality assessment.

Bartneck et al. (2021) discuss AI (Artificial Intelligence) applications in the military domain, specifically, aspects such as autonomy and control regarding the use of autonomous weapon systems. On this, the authors stress that the outcome of targeting the Malaysia Airlines MH17 jet (from Amsterdam to Kuala Lumpur in 2014) would be different if an advanced system would have been put in place for making a clear distinction between civilian and military aircrafts. Moreover, the same authors highlight the multi-source nature of the information received when conducting proportionality assessment. Remaining in the same area of autonomous weapons, Morgan et al. (2020) assess that the principle of proportionality implies a case-by-case assessment which is a subjective process in which 'the harm of possible collateral effects is weighted against the importance of the military objective', while further concluding that the proportionality assessment is "an evaluative, qualitative, and ethical assessment by a human weighting and comparing complex values". Along these lines, Thome (2020), Holland (2002) and Katzir (2018) argue that proportionality assessment relies on "less concrete concepts" like human judgement, values, and foresight while requiring a certain degree of military operational expertise. On the other side, Whittemore (2015) and Kai (2022) consider that heuristics and different types of biases are also involved in this process. In the naval domain, Lee (2021) conducts an in-depth analysis on the role and participation of civilians in the execution of military operations, and further calls for a broader perspective and a larger dialogue on interpreting and applying the principle of proportionality. In the air domain, Gul (2021) discusses the applicability of the principle of proportionality to drone attacks and addresses a topic scarcely tackled in the existing body of knowledge, i.e., military advantage which plays a direct and concrete role as one of the two key components of the principle of proportionality by directly contributing to the achievement of the military objectives defined. Going to the military cyber domain, Pascucci (2017) analyzes what the principles of distinction and the principle of proportionality mean. On this behalf, the author considers that the principle of proportionality is an *ex-ante* analysis that conducts to the *ex post* facto consequences. These consequences are not relevant when determining if the principle of proportionality was violated since the principle of proportionality is governed by uncertainty and is conducted before target engagement. Moreover, In the military cyber domain, Maathuis, Pieters & van den Berg, (2021) consider as influential factors concepts like military aim, the background, and experience of the military decision maker (i.e., military Commander) as well as the willingness/appetite and stress.

3. Research Methodology

To identify and capture the influential factors contributing to proportionality assessment decisions and be able to further model them, the following research questions are formulated:

- What are the influential factors contributing to proportionality decisions?
- How to model the influential factors to support proportionality decisions?

These research questions imply identifying and embedding multiple dimensions existing in different domains, i.e., (cyber) military operations, law, behavioural psychology, and AI. Hence, a call for multidisciplinary research considering a socio-technical perspective by means of building a model as an artefact with direct societal meaning, contribution, and impact. Accordingly, a Design Science Research methodological approach is considered following the succeeding research activities (Peffers et al., 2007; Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018):

Problem Definition and Aim: old as well as modern and recent military operations demonstrate their complex nature and implications plus the broad range of influential factors contributing to proportionality decisions, i.e., a key element and decisive moment in their execution. While different aspects and dimensions have been defined and analysed on this behalf, the existing need for gathering and understanding the influential factors to such decisions continues to grow as the means and methods of conducting war are becoming more intelligent and adaptive to different (new) situations. On this behalf, this research captures influential factors to proportionality decisions through extensive literature review conducted using combinations of keywords like 'proportionality', 'targeting', 'assessment', 'test', and 'influence' in scientific databases such as IEEE Xplore, ACM Digital Library, Springer Link, and Google Scholar.

Design and Development: once the influential factors are gathered, they are analysed and clustered based on their nature, and further captured in an intelligent model using BBN.

Evaluation and Dissemination: afterwards the proposed model is evaluated through demonstration considering

three different situations of influence on proportionality decisions. Furthermore, the results obtained are interpreted and positioned in their corresponding situations through these article and upcoming scientific presentations.

4. Influential Factors

With respect to our research goal, to structure the influential factors identified, a clustering approach is considered based on (Da Veiga et al. 2020; Maathuis, Pieters & van den Berg, 2021; Maathuis & Chockalingam, 2022) meaning military context, military-legal, and human clusters of variables. Correspondingly, the clusters are depicted in Figure 1, below addressed, and structured in Table 1.

- Military Context: this cluster contains variables characterizing different practical aspects regarding the strategic, operational, and tactical levels of war.
- Military-legal: this cluster embeds variables related to specific military-legal dimensions applicable when building, executing, and assessing military operations in respect to the appropriate legal frameworks.
- Human cluster: this cluster immerses variables linked to human aspects such as behavior and values.

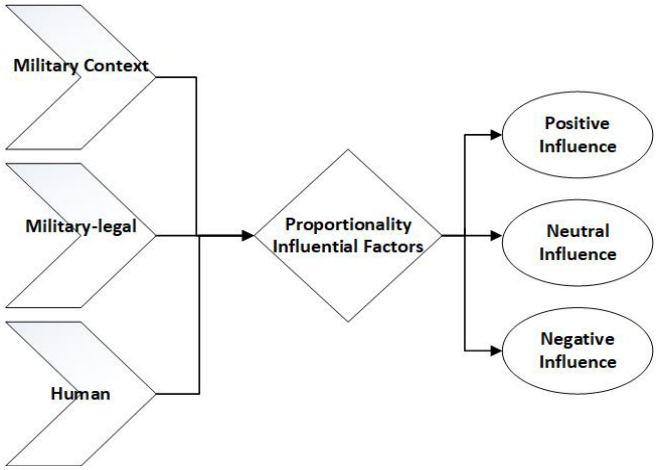


Figure 1: Proportionality decisions influential factors

Table 1: Proportionality decisions influential factors

Cluster	Variables and States	References
Military Context	Military setting: SIMPLE, COMPLEX Military goal/purpose: ACHIEVABLE, NON-ACHIEVABLE. Military requirements and conditions: UNCLEAR/ILL-DEFINED, CLEAR/WELL-DEFINED Lessons learned from previous operations: PRESENT, NOT PRESENT Military intelligence: POOR, ADVANCED	Katzir (2018); Maathuis, Pieters & van den Berg, (2021); Gul (2021); Hayashi (2010); Bartneck et al. (2021); Morgan et al. (2020); McKenna (2020).
Military-legal	Compliance with military-legal principles: LOW, MEDIUM, TOTAL {Compliance with/Status} ROEs: UNCLEAR, CLEAR	US Army (2013); NATO (2016); Pascucci (2017); Colonomos (2017); Katzir (2018); Ali (2020); Thome (2020); Maathuis, Pieters & van den Berg, (2021); Bartneck et al. (2021).
Human	Background: LOW, HIGH Training and Experience: JUNIOR, SENIOR Willingness/risk appetite: LOW, HIGH Stress: LOW, HIGH Broader effects perspective: ENABLED, DISABLED Foresight capability: LOW, HIGH Religion, Culture, Values, Moral identity, Moral disengagement, Personality type (e.g., authoritarian), Depression, Drinking problems, Genre, Bias (e.g., overconfidence, confirmation)	Holland (2002); Maathuis, Pieters & van den Berg, (2021); Gul (2021); Henderson, I., & Reece, (2018); Hayashi (2010); Fenrick (2001); Whittemore (2015); Kimhi & Kasher (2015); Colonomos (2017); de Graaff, Giebels, & Verweij, (2020); Morgan et al. (2020); Thome (2020); LeardMann et al. (2013); Kai (2022).

5. Model Design

To model the influential factors identified while taking into consideration the uncertainty and subjectivity that govern the proportionality assessment process as well as the complexity surrounding it (Pourret, Na & Marco, 2008), a BBN modelling approach is considered for designing and developing the model proposed in this research. This approach already proved its applicability in modelling complex systems and processes within diverse plethora of societal domains (Marcot & Penman, 2019), and contains two main components: (i) *qualitative* which is a Directed Acyclic Graph (DAG) as depicted in Figure 2 that captures a set of variables and direct cause-effect relationships (i.e., edges) between them; and (ii) *quantitative* component that contains the Conditional Probability Tables (CPTs) with the conditional probabilities for all possible combinations of child-parent variable states, as shown in Figure 2. It is important to note that in case a variable does not have a parent variable, the CPT contains the priori marginal probabilities of the corresponding variable. Furthermore, with respect to the existing reasoning types of BBN (i.e., predictive, diagnostic, intercausal, and combined), our model embeds the predictive reasoning approach as it is reasoning from cause (upper layer) to effect (lower layer) (Maathuis & Chockalingam, 2022).

The clusters of our proposed model together with their corresponding variables and states are captured in Table 1. The variables *religion, culture, values, moral identity, moral disengagement, personality type, depression, drinking problems, bias,* and *genre* were not included in the model as they are indirect influential factors, are difficult to assess, and are less tackled in the scientific literature in relation to the proportionality assessment. For these factors, more research is necessary, thus this is a future research line for extending the present model. The target variable (proportionality influence) contains the following values: Negative, Neutral, and Positive.

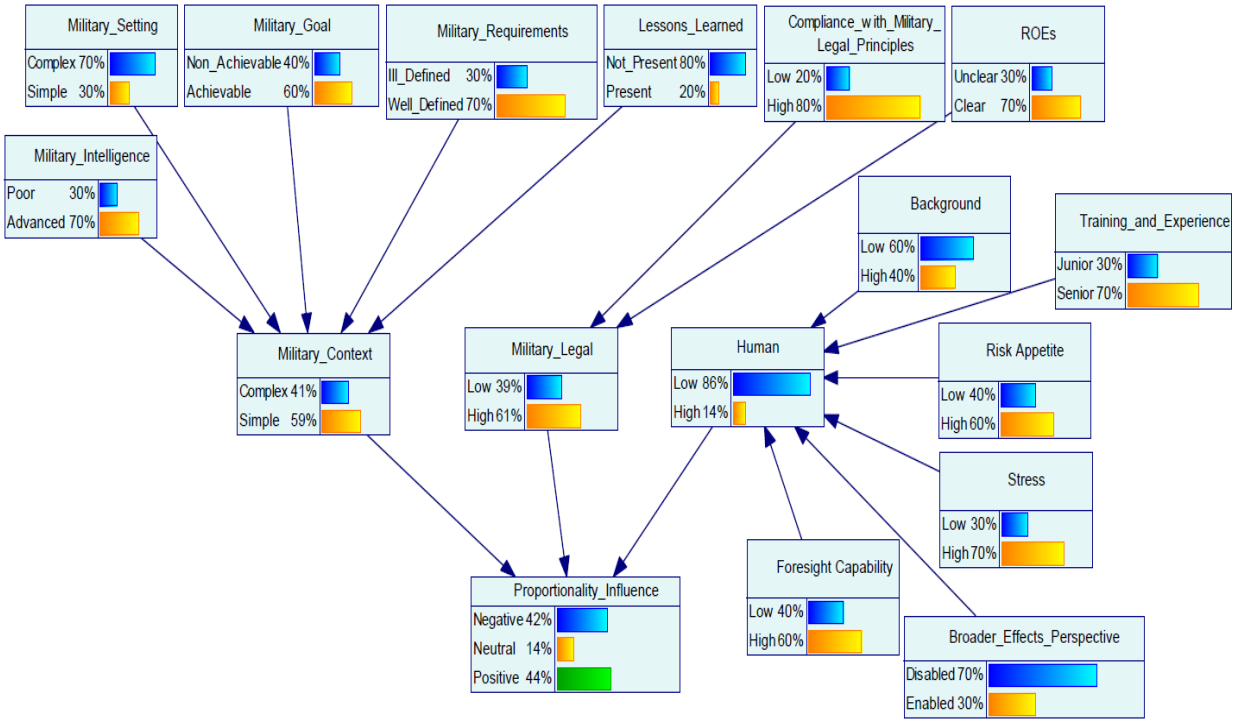


Figure 2: Proposed model without evidence provided

The conditional probabilities used in the proposed model are example values. This is to demonstrate the use of such model by providing evidence to different variables in illustrative case scenarios in Section 6. Typically, such probabilities can be elicited from experts and empirical data. In addition, approaches like Noisy-OR (Iemmer & Gossink, 2004) can be used to reduce the number of conditional probabilities to elicit from experts and complete the CPTs.

6. Model Evaluation

As discussed in Section 3, the evaluation of this model is conducted through demonstration on three different case scenarios.

In the first scenario, evidence is set in the upper layer: Military Setting = “Complex”, Military Requirements =

"Ill-defined", and RoEs = "Unclear", Training and Experience = "Junior", and Stress = "High". Then, the posterior probability is calculated by the model for its other variables without any evidence. Hence, as depicted in Figure 3, the results influence is "Negative", meaning that the Proportionality influence has a 57% chance of being Negative, 13% chance of being Neutral, and 30% chance of being Positive. This could be the case of a military Commander that is less experienced and has to decide when dealing with major amount of stress and unclarities regarding the military and legal requirements and possibilities.

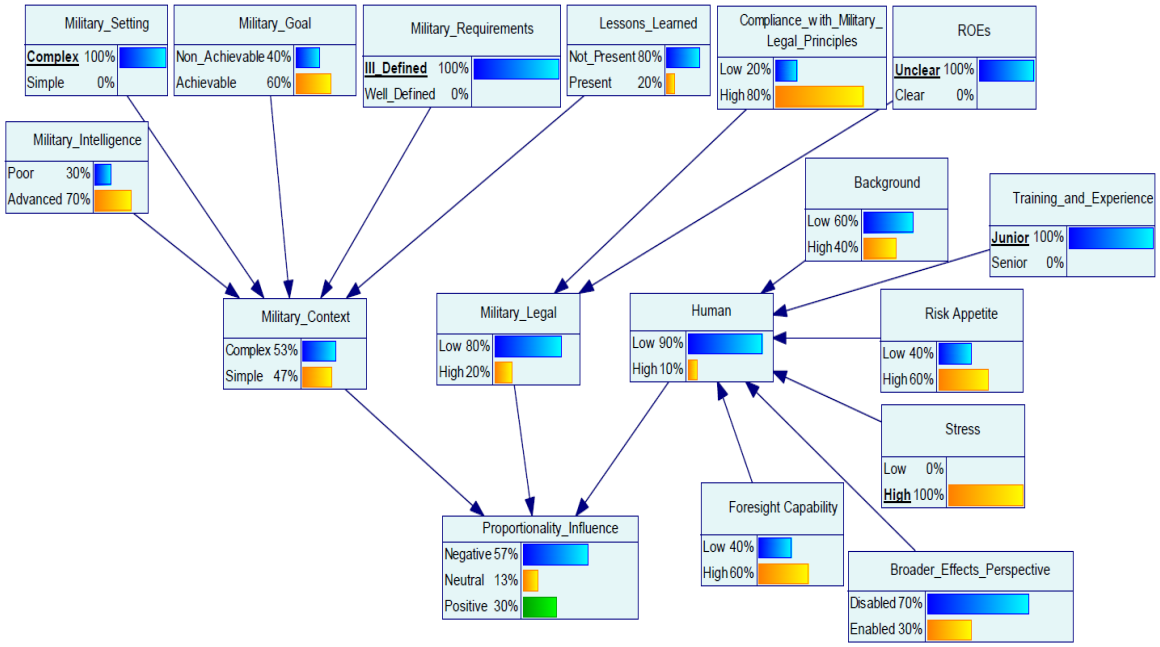


Figure 3: First case scenario

In the second scenario, evidence is set in the upper layer: Military Goal = "Achievable", Military Requirements = "Well-defined", RoEs = "Clear", Training and Experience = "Senior", and Stress = "High". Then, the posterior probability is calculated by the model for its other variables without any evidence. Hence, as shown in Figure 4, the influence is "Positive", meaning that the Proportionality influence has a 49% chance of being Positive, 14% chance of being Neutral, and 36% chance of being Negative**.** This could be the case of a military Commander that is experienced, which must make this decision decision not dealing with high stress while the military and legal requirements and conditions are clear.

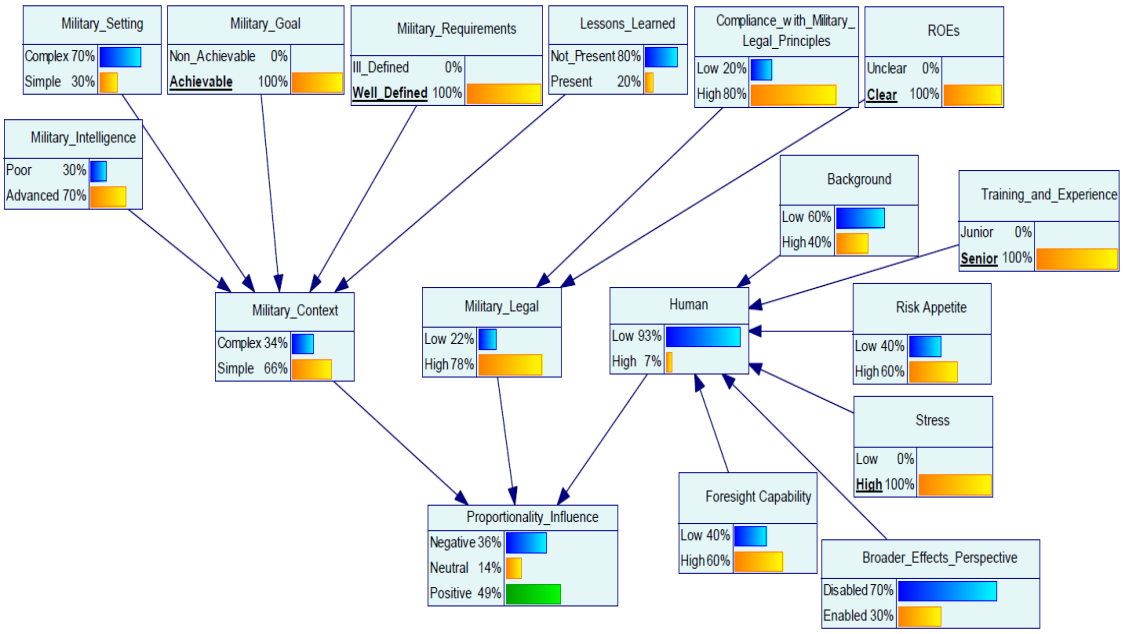


Figure 4: Second case scenario

In the third scenario, evidence is set in the upper layer: Military Intelligence = "Poor", Lessons Learned = "Not_Present", Training_and_Experience = "Junior", Risk Appetite = "High", Stress = "High", and Broader_Effects_Perspective = "Disabled". Based on this evidence, the posterior probability is calculated for the remaining variables. Conclusively, as Figure 5 illustrates, the influence is "Negative". This case could correspond to the situation when a military Commander is less experienced, does not benefit from previous lessons learned (i.e., knowledge gathered from an antecedent military operation), and deals with a high amount of stress when making this decision.

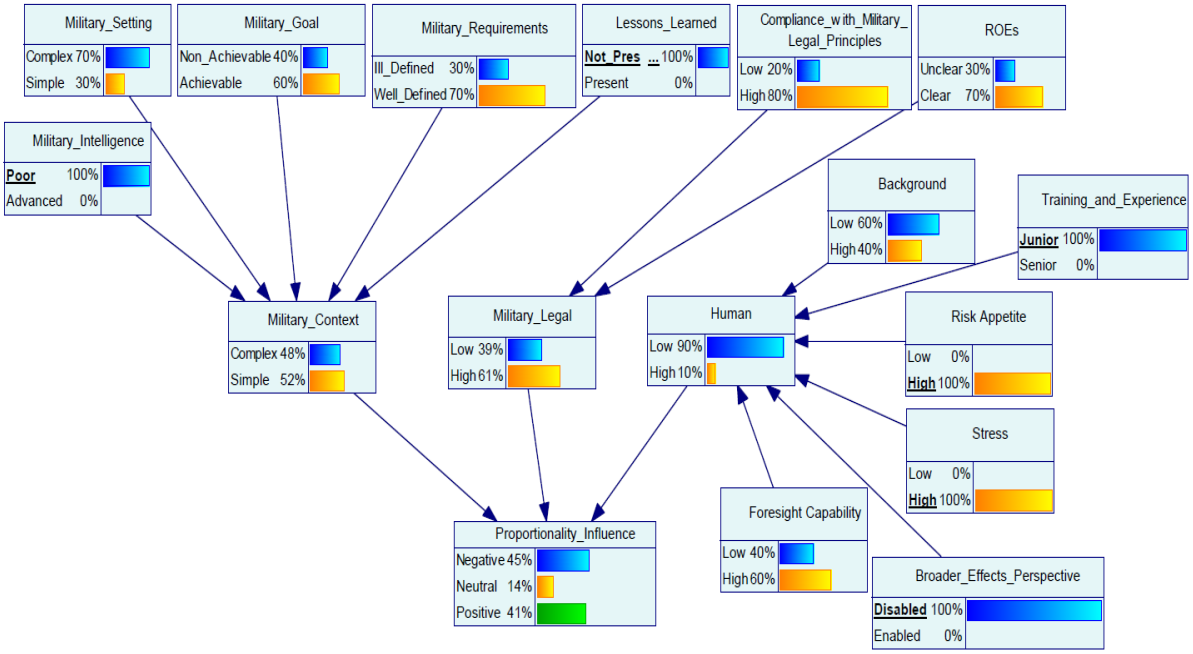


Figure 5: Third case scenario

7. Conclusions

To all intends and purposes, the laws of war could be reduced to a "subtle equilibrium between two diametrically opposed impulses: military necessity and humanitarian considerations" (Dinstein, 2016). Specifically, through its principle of proportionality and its corresponding proportionality assessment bring together two distinct military-legal constructs that should be continuously assessed on the future version of reality through the current version of reality taking into consideration the information herein available. This is a difficult process and a challenging task (Katzir, 2018) taking into consideration the high level of complexity and uncertainty surrounding the two key components of this assessment, i.e., collateral damage and military advantage together with the interpretation of the key operation between them, i.e., the fact that the expected collateral damage should not be excessive in relation to the anticipated collateral damage. All in all, as Colonomos (2017) stresses, proportionality is born as a "compromise between political constraints, military necessities, legal claims, and ethical aspirations".

Taking this into account, significant scientific and practitioner attention is provided for analyzing and applying these perspectives, but limited attention is dedicated to understanding other factors that play a role in influencing the reasoning process of the military decision-makers that conduct the proportionality assessment. In this respect, the present research investigates and clusters influential factors participating when the proportionality assessment is conducted and further capture and model them by means of a BBN model designed pursuing a Design Science Research methodological approach. Finally, the model proposed is demonstrated through three case scenarios for understanding the influence variation of the identified factors. Through this model, this research intends to bring awareness to military and political decision-makers, and to encourage further research on identifying, understanding, and modelling other influential factors to proportionality assessment in diverse missions and operational scenarios that would contribute to both human and machine transparent and responsible behavior.

References

Additional Protocol I (1977). Art 51(5)(b) – Protection of the civilian population.

Ali, S. (2020). Coming to a Battlefield Near You: Quantum Computing, Artificial Intelligence, & Machine Learning's Impact on Proportionality. *Santa Clara J. Int'l L.*, *18*, 1.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). Military Uses of AI. In *An Introduction to Ethics in Robotics and AI* (pp. 93-99). Springer, Cham.

Cannizzaro, E. (2006). Contextualizing proportionality: jus ad bellum and jus in bello in the Lebanese war. *International Review of the Red Cross*, *88*(864), 779-792.

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 26-35).

Colonomos, A. (2017). Proportionality in Warfare as a Political Norm. *Weighing Lives in Warfare*, 217-240.

de Graaff, M. C., Giebels, E., & Verweij, D. E. (2020). On moral grounds: Moral identity and moral disengagement in relation to military deployment. *Military Psychology*, *32*(4), 363-375.

Dinstein, Y. (2016). The conduct of hostilities under the law of international armed conflict. Cambridge University Press.

Fenrick, W. J. (2001). Targeting and proportionality during the NATO bombing campaign against Yugoslavia. *European Journal of International Law*, *12*(3), 489-502.

Gall, L. (2004). International Criminal Tribunal for the Former Yugoslavia (ICTY): Prosecutor v. Galic. *International Legal Materials*, *43*(4), 789-879.

Gillard, E. C. (2018). Proportionality in the conduct of hostilities: the incidental harm side of the assessment.

Gül, Y. E. (2021). Drone Attacks and the Principle of Proportionality in the Law of Armed Conflict. In *Annales de la Faculté de Droit d'Istanbul* (No. 70, pp. 119-145).

Hayashi, N. (2010). Requirements of military necessity in international humanitarian law and international criminal law. *BU Int'l LJ*, *28*, 39.

Henderson, I., & Reece, K. (2018). Proportionality under International Humanitarian Law: The Reasonable Military Commander Standard and Reverberating Effects. *Vand. J. Transnat'l L.*, *51*, 835.

Holland, J. C. (2002). *Military Objective and Collateral Damage: Their Dynamics and Relationship*. JUDGE ADVOCATE GENERAL'S SCHOOL CHARLOTTESVILLE VA.

Hawkins, J. A. (2003). Joint Tactics, Techniques, and Procedures for Intelligence Support to Targeting. USAF.

Kai, C. C. J. (2022). Cognitive Biases: The Root of Irrationality in Military Decision-Making. *Pointer, Journal of the Singapore Armed Forces*, *42*(2).

Katzir, R. (2018). Four Comments on the Application of Proportionality under the Law of Armed Conflict. *Vand. J. Transnat'l L.*, *51*, 857.

Kimhi, S., & Kasher, A. (2015). Moral dilemmas in military situations: Proportionality principle, religiosity, political attitudes, and authoritarian personality. *Military Psychology*, *27*(3), 169-184.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Lee, D. T. (2021). The Principle of Proportionality in Maritime Armed Conflict: A Comparative Analysis of the Law of Naval Warfare and Modern International Humanitarian Law. *Sw. J. Int'l L.*, *27*, 119.

LeardMann, C. A., Powell, T. M., Smith, T. C., Bell, M. R., Smith, B., Boyko, E. J., ... & Hoge, C. W. (2013). Risk factors associated with suicide in current and former US military personnel. *Jama*, *310*(5), 496-506.

Lemmer, J. F., & Gossink, D. E. (2004). Recursive noisy OR-a rule for estimating complex probabilistic interactions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(6), 2252-2261.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2016). Cyber weapons: a profiling framework. In *2016 International Conference on Cyber Conflict (CyCon US)* (pp. 1-8). IEEE.

Maathuis, C., Pieters, W., & Van den Berg, J. (2018). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (pp. 1-6). IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2021). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*, *17*(2), 352-374.

Maathuis, C., & Chockalingam, S. (2022). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security*, 21(1).

Maathuis, C. (2022a). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1, pp. 170-177).

Maathuis, C. (2022b). On Explainable AI Solutions for Targeting in Cyber Military Operations. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 166-175).

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental modelling & software*, *111*, 386-393.

McKenna, M. B. (2020). The Discourse of Proportionality and the Use of Force: International Law and the Power of Definition. *Nordic Journal of International Law*, *89*(3-4), 364-382.)

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military applications of artificial intelligence: ethical concerns in an uncertain world*. RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA United States.

NATO (2016). AJP-3.3 Allied Joint Doctrine for Air and Space Operations. *Edition B Version*, *1*.

Oxford Institute for Ethics, Law, and Armed Conflict (2009). Proportionality in War. Workshop Report.
Pascucci, P. (2017). Distinction and Proportionality in Cyberwar: Virtual Problems with a Real Solution. *Minn. J. Int'l L.*, *26*, 419.
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.
Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.
Pourret, O., Na, P., & Marcot, B. (Eds.). (2008). *Bayesian networks: a practical guide to applications*. John Wiley & Sons.
U.S. (2013). JP-3.60 Joint Targeting. Joint Publication.
Preston, S. E., & Taylor, R. S. (2016). *Department of Defense Law of War Manual*. General Counsel of the Department of Defense Washington United States.)
Whittemore, L. A. (2015). Proportionality Decision Making in Targeting: Heuristics, Cognitive Biases, and the Law. *Harv. Nat'l Sec. J.*, *7*, 577.

Social Media Manipulation Awareness through Deep Learning based Disinformation Generation

Clara Maathuis[1] and Iddo Kerkhof[2]
[1]Open University of the Netherlands, Heerlen, The Netherlands.
[2]NLP Software Engineer, The Netherlands.

clara.maathuis@ou.nl
email@iddo.eu

Abstract: As a digital environment introduced for establishing and enhancing human communication through different social networks and channels, social media continued to develop and spread at an incredible rate making it difficult to find or imagine a concept, technology, or business that does not have or plan to have its social media representation and space. Concurrently, social media became a playground and even a battlefield where different ideas carrying out diverse validity degrees are spread for reaching their target audiences generated by clear and trustable well-known, uncertain, or even evil aimed entities. In the stride carried out for preventing, containing, and limiting the effects of social manipulation of the last two types of entities, proper/effective security awareness is critical and mandatory in the first place. On this behalf, several strategies, policies, methods, and technologies were proposed by research and practitioner communities, but such initiatives take mostly a defender perspective, and this is not enough in cyberspace where the offender is in advantage in attack. Therefore, this research aims to produce social media manipulation security awareness taking the offender stance by generating and analysing disinformation tweets using deep learning. To reach this goal, a Design Science Research methodology is followed in a Data Science approach, and the results obtained are analysed and positioned in the ongoing discourses showing the effectiveness of such approach and its role in building future social media manipulation detection solutions. This research also intends to contribute to the design of further transparent and responsible modelling and gaming solutions for building/enhancing social manipulation awareness and the definition of realistic cyber/information operations scenarios dedicated/engaging large multi-domain (non)expert audiences.

Keywords: information operations, cyber operations, social manipulation, disinformation, misinformation, security awareness, machine learning, deep learning.

1. Introduction

"The success of manipulation depends on the level of conviction and force of the denial." (Tess Binder)

It would be difficult to recall a modern or current societal and technological trend, topic, or event that is not projected in the digital realm through social media discourses and not surrounded by diverse manipulation mechanisms like disinformation and misinformation (Maathuis & Chockalingam, 2022b). In this realm, agents, e.g., public opinion organizations, independent bodies, and civilians carry out different activities and engage (un)intentionally with diverse manipulation forms which impact their behaviour individually and collectively (Bastick,2021; Chockalingam & Maathuis, 2022). Engaging the target audience through social manipulation can be done through vectors, e.g., user profiling being (i) demographic meaning identifying individuals' unique characteristics, beliefs, needs, and vulnerabilities, and (ii) psychometric implying personality-based segmentation, and micro-targeting based on analyzing/altering audience's personal actions (Kertysova, 2018; Fard & Maathuis, 2021). While efforts dedicated to limiting, controlling, and preventing social manipulation using AI-based techniques, e.g., Twitter and Facebook relying on Machine Learning-based solutions for stamping out trolls, finding and removing fake bot accounts, and proactively identifying sensitive content (Perez-Escolar et al. 2021), still these mechanisms are insufficient as the attackers are intelligent in developing adaptive techniques that succeed on bypassing defending mechanisms and reach their goals. Hence, changing the paradigm and treating this phenomenon through the eyes of the attacker while considering building datasets and transparent intelligent solutions (Maathuis, 2022b) that model and draw relevant principles and requirements for building awareness and defending techniques could be (the basis of) a solution in this direction. Hence, this research aims to build social manipulation security awareness through a deep learning model for generating disinformation tweets. To achieve this objective, multidisciplinary approach is conducted by merging studies from social media, cyber security, information/cyber operations, and deep learning domains through a Design Science Research methodology following a Data Science approach having the following contributions:

- Opening a path to scientific and practitioner communities taking offender's stance when building/enhancing existing social manipulation security awareness through content generation.

- Encouraging increasing the variety of social media security awareness solutions that are transparent and responsible being set up by security, social media, and AI experts dedicated to multiple audiences by integrating various disciplines/visions.
- Supporting decision-making processes of policy, security, and social media decision-makers when building strategies and policies for preventing, containing, and/or limiting social manipulation through human and technological-based solutions.

The remainder of this article is structured as follows. In Section 2, the background and related studies of this research are discussed. In Section 3, the research approach taken is elaborated while presenting research activities taken. In Section 4, the deep learning techniques used to build the model are tackled. In Section 5, the implementation process is depicted together with the results obtained. In Section 7, concluding remarks on the findings obtained and future research ideas are addressed.

2. Social Manipulation Research

In this section the multidisciplinary nature of this research is expressed by discussing studies, methods, and strategies relevant to achieving the aim of this research. The EU Commission stresses the importance of protecting values and democratic systems by tackling social manipulation mechanisms like disinformation, defined as "false or misleading content that is spread with an intention to deceive or secure economic or political gain" (EU Commission, 2022a). Among the initiatives established are the Action plan on disinformation, the European Democracy Action Plan, the Covid-19 disinformation monitoring programme, and the Strengthened Code of Practice on Disinformation which calls for values like transparency of political advertising, empowering researchers (e.g., through public data availability), and empowering the fact-checking community (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b). For increasing digital resilience and combating disinformation for achieving ideological, political, or financial goals, NATO acknowledges the damage done to "citizens' faith in the institutions of democratic governments and resources of public information and discussion" (NATO, 2021a) and calls for developing fact-based solutions and credible public communications based on (i) understanding and analysing the information environment, (ii) engaging audiences to build resilience, and (iii) communicating proactively and exposing major disinformation cases (NATO, 2021b). As such solutions are currently implemented integrating AI, according to Rathenau (2022), they can help managing credibility; however, can also amplify conspiracy.

It is then important to take a step back and see the whole context. Accordingly, Ki-Aries & Faily (2017) argue for designing user-specific security awareness solutions by incorporating human factors in design through a series of targeted and actionable topics that consider persona characteristics, i.e., (i) goal directed perspective, (ii) role-based approach, (iii) fiction-based perspective through human intuition and assumptions, and (iv) engaging perspective through data use. Reisach (2021) analyzes the role and responsibility that social platforms have on conducting diverse businesses, expressing political discourses and election decisions, integrating AI techniques for different functionalities, and related to the formation, execution, and impact produced by different manipulation incidents. Moreover, the author argues that responsibility is shared and global – of society, social platforms, and users. To support it in a proactive way, the author argues for building education solutions oriented on (i) tech and media literacy, (ii) educational support for vulnerable populations, and (iii) education and awareness to foster an ethical mindset. Chen, Chen & Xia (2022) analyze user's behavior in social networks arguing that social media is becoming an environment for cyber weapons deployment (Maathuis, Pieters & van den Berg, 2016) and a force multiplier in warfare through goals that include public opinion reconnaissance, sentiment analysis, and active intervention. The authors reflect on the actors involved, nature of the environment (warfare versus infosphere), and the arsenal used, e.g., bot, botnet, troll, cyborg, plus behavior information representation, user recognition, and homogeneous community discovery. Shu (2022) conducts a survey on recent advancements and stresses the existing data challenge, i.e., availability of public datasets useful for understanding manipulation styles/methods. While the author acknowledges the significant use of deep learning, the argument for building not just proactive, accurate, and robust disinformation countering solutions is put forward, and as important, that the solutions should be trustworthy, i.e., investigating how to (i) transform meaningful human cognition into knowledge, (ii) incorporate noisy, incomplete, and complex human feedback for better representation learning, and (iii) interpret prediction results with knowledge reasoning and causal discovery. An experiment was conducted by Bastick (2021) with 233 undergraduate students for investigating behavioral effects of fake news. The results reveal that fake news have the real potential of altering individuals' behavior and cluster disinformation effects as empirically observed, refined, and predicted by the malicious agents producing them. Moreover, the author calls for urgent development of solutions that generate

experiments for strengthening users' protection from manipulation mechanisms. Perez-Escolar et al. (2021) conduct a study with 150 students for strengthening fact-checking skills and project-based learning, and mention formulating judgements based on incomplete or limited information as a major competence.

On AI applications, Kertysova (2018) stresses that they are double-edged since they provide useful tools for building powerful and scalable solutions for countering social manipulation while they imply own set of limitations and unintended consequences. Additionally, AI is the backbone since through its techniques facilitates the development and spreading of social manipulation content to large audiences. Particularly, the model of Kula, Kozik & Choras (2021) is based on a BERT architecture named RoBERTa for dealing with Covid-19 fake news using a mix of datasets. The results obtained are promising (accuracy above 90%) and reflect the potential of such techniques. In the survey by Wang et al. (2022), the authors propose as further research developing proper interpretation methods, enhancing robustness to possible security and privacy attacks, properly handing with imbalanced data, and merging GAN with other deep learning techniques. An example of successfully combining GAN with BERT techniques is developed by Yu et al. (2020) for representing medical concepts and discovering annotation inconsistency on medical labels. Moreover, Shu, Li, Ding & Liu (2021) generate synthetic news through the FACTGEN (Fact-Enhanced Synthetic News Generation) method which implements a PSA (Pseudo-Self-Attentive) language model that tackles the challenges on generating news content related to a given claim and ensuring that the generated content contains supplemental fact information. For generating real facts through (journalistic) news, Leppanen et al. (2017) consider that the following criteria should be respected: transparency, accuracy, system modifiability and transferability, output fluency, data availability, and news topicality. These criteria are considered when building the artefact proposed in this research. However, strengthening social media security spanning cyber incidents' life cycle implies building not only AI-based security awareness solutions for educational purposes, but also building intelligent solutions for testing attack resistance against possible cyber attacks. Le, Wang & Lee (2020) generate malicious content for attacking existing security mechanisms embedded in deep learning-based fake news detection models. The solution proposed is based on GAN and contains the attack module, conditional text generator, and style module. Nevertheless, a limited number of disinformation datasets are available and the necessity of proposing new ones is experienced by researchers and practitioners in this domain. Accordingly, Kim et al. (2021) propose the FibVID (Fake news information-broadcasting dataset of Covid-19) with tweets containing Covid-19 and non-Covid-19 fake news for analysis and diffusion of fake news. Another dataset is advanced by Patwa et al. (2021) focusing on Twitter with data labelled as real and fake used for building baseline ML models.

As this extensive literature study reflects, the harm produced by disinformation is experienced by society, by social platforms, and their users. While different incentives and mechanisms for proactively and reactively tackling disinformation incidents exist, social media security awareness to such manipulation mechanisms is still in its infancy due to issues like limited data availability, lack of multidisciplinary/interdisciplinary/transdisciplinary/anti-disciplinary expertise, fast technological advancements, and the existing (un)known vulnerabilities of humans and social platforms. This is the motivation behind our research.

3. Research Methodology

Aiming at building an intelligent model that generates social manipulation disinformation content in a transparent and responsible way for building/enhancing social media security awareness of various audiences, in this research the following research questions are formulated:

- How to build a deep learning model for generating social manipulation disinformation content to support social media security awareness?
- What are the results obtained based on evaluating the model proposed?

To answer these questions and reach the goal formulated, the Design Science Research methodology is followed in a Data Science approach based on multidisciplinary research (Venable, Pries-Heje & Baskerville, 2017; Maathuis, Pieters & van den Berg, 2018a) considering the design requirements for user-oriented security awareness requirements established by Ki-Aries & Faily (2017). Accordingly, a deep learning model is developed and evaluated through the following research activities (Peffers, 2018):

Problem definition and solution aim: the significant increase in number, diversity, and scale of effects of social manipulation incidents shows that social media platforms, users, and governments are at the beginning of a long path concerning proactive security awareness and dealing with such incidents. Moreover, the ongoing Covid-19

pandemic and war in Ukraine are proofs that such incidents do not only influence users believes but also their behaviour towards the topics involved and (in)direct relational concepts, e.g., natural occurrence of the Covid-19 virus and its realistic effects in relation to other viruses. Hence, given existing efforts that mainly adopt defender's vision, in this research offender's vision is taken for building a deep learning model that generates disinformation content for building/enhancing security awareness. Accordingly, multidisciplinary research is conducted starting with extensive literature review on existing research and practitioner strategies, policies, and methods using different combination of keywords like 'social media', 'manipulation', 'disinformation', 'awareness', 'generation', and 'deep learning' in scientific databased like IEEE Digital Library, ACM Digital Library, Scopus, and Google Scholar.

Solution development: to build the model, data is collected surrounding the Covid-19 pandemic tweets concerning the first two waves, i.e., first wave between 01.01.2020 and 21.03.2020 and second wave between 15.10.2020 and 31.12.2020. Accordingly, the data collection, processing, hashtag counting, topic extraction, model implementation using BERT and GAN deep learning techniques, and evaluation processes are implemented in Python, are compliant with the criteria of Leppanen et al. (2017), and are further discussed. Accordingly, the number of mentions per hashtag is depicted in Figure 1 below:

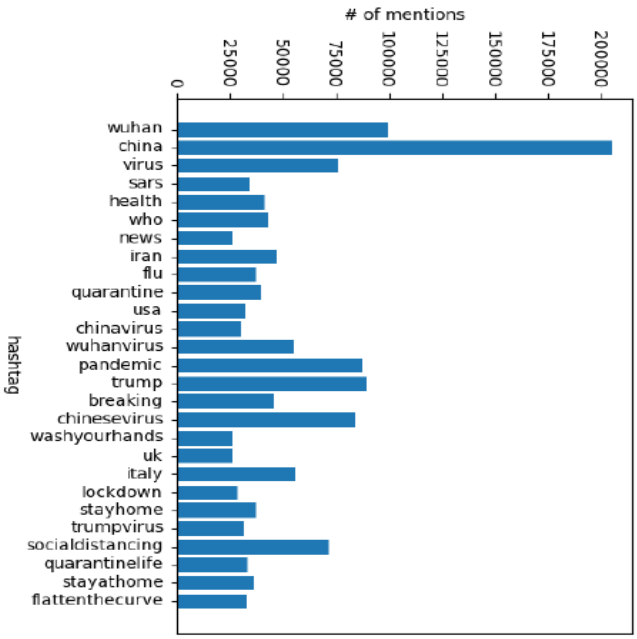


Figure 1: Number of mentions per hashtag

Solution evaluation: the evaluation of the model proposed is carried out continuously through technical and human-based assessment for examining text quality of the content generated and further improving through optimization the solution proposed.

Communication: the approach followed, and the results obtained are herein presented emphasizing the real need for datasets publicly available and arguing for adopting offender's stance when building intelligent solutions for enhancing/building social manipulation security awareness.

4. Modelling Techniques

Nowadays is difficult to point to a field, problem, or system not using or would not benefit from using AI (Maathuis, 2022a). At the core of its current development is Machine Learning which deals with learning from experience through different learning paradigms, e.g., in a predictive approach by mapping input to output given labelled input-output pairs (supervised learning) or in a descriptive approach through knowledge discovery where only the input is provided for learning patterns from data (unsupervised learning) (Murphy, 2012). Due to technological advancements and data availability characterizing, while stepping out of the AI winters, a particular area of ML that amazes through applicability is deep learning (Janiesch, Zschech & Heinrich, 2021). In deep learning are used (complex) neural architectures containing different layers, i.e., input, (one/multiple) hidden, and output where through different operations, the output is computed (Goodfellow, Bengio & Courville, 2016). As this research uses specific deep learning techniques (BERT-Bidirectional Encoder

Representations from Transformers and GAN-Generative Adversarial Networks), they are further discussed.

GAN is a deep learning-based generative technique proposed by Goodfellow et al. (2020) where two distinct neural networks named generator and discriminator built for competing in a game where one should win and the other should lose as in Figure 2. The generator is generating plausible data which should appear real, and the discriminator learns to distinguish between fake data produced by the generator from real data. Hence, the output produced by the generator is fed into the discriminator and is considered a set of negative training data examples. Moreover, the discriminator learns through backpropagation of the loss compared to a fake classification, while the generator learns by comparing the loss to a real classification since it tries to trick the discriminator (Creswell et al., 2018; Aggarwal, Mittal & Battineni, 2021). The training stops when either the generator or the discriminator stops improving, or when the discriminator cannot distinguish fake from real data anymore.

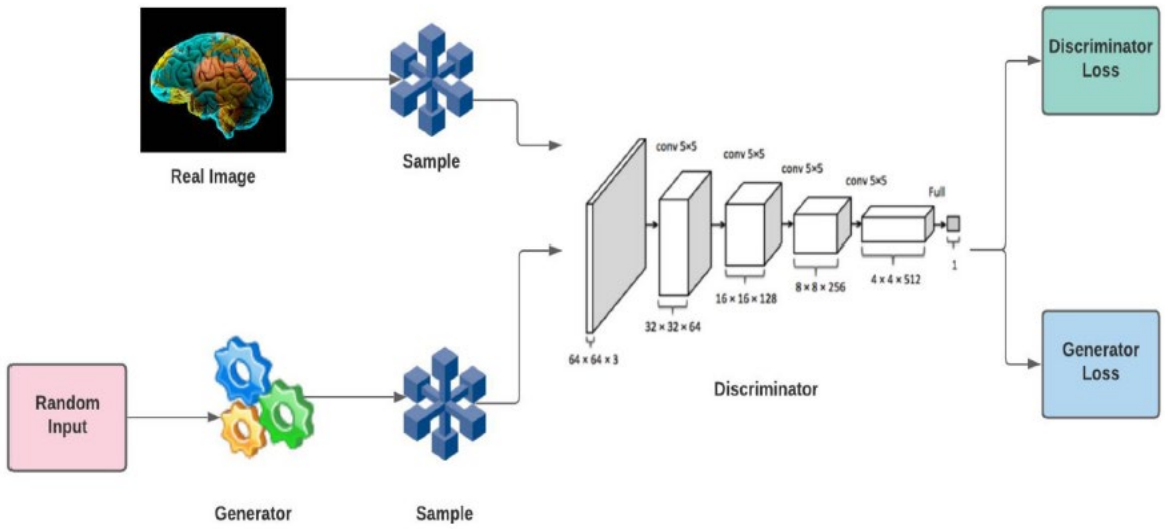


Figure 2. GAN block diagram from (Aggarwal, Mittal & Battineni, 2021)

BERT is a deep learning transformer-based technique proposed by Delvin et al. (2018) at Google for enhancing contextual understanding of unlabelled text across a wide range of activities. BERT is the first deeply bidirectional (left-to-right and right-to-left) unsupervised language architecture containing twelve layers where with each layer more contextual information is learned, see Figure 3. Specifically, the input layer corresponds to tokenized training data and the output layer to the masked language model for which it is trained while aiming to predict the original value of the masked tokens considering the context from both directions (Delvin & Chang, 2018; Qiu et al. 2020) which makes it suitable for text embeddings and text classification tasks.

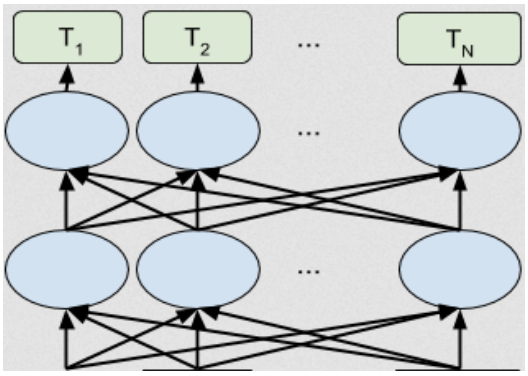


Figure 3: BERT architecture from (Delvin & Chang, 2018)

5. Model Implementation and Results

While datasets useful for tackling disinformation continue to be made available, currently their number is still limited. Hence, for building a proactive approach by expanding the space of available datasets, a merge between GAN and BERT is considered based on previous promising results shown in different domains, e.g., pharmaceutical text classification (Auti et al., 2022), aggressive and violent incidents identification in social

media (Ta et al., 2022), and stock price prediction (Sonkiya, Bajpai & Balsal, 2021). This process is conducted in two steps: generating data without considering whether they imply disinformation or not and generating disinformation data or transforming regular data into disinformation data. While merging these techniques, the hypothesis is that the language understanding of the pretrained BERT model will be transferred to GAN which leads to a better text generation than when only using BERT. Accordingly, the Huggingface transformers library (Wolf et a. 2019) is used, and three experiments are conducted.

In the first experiment, a pre-trained BERT model is used to fill masked words in tweets. When a single word is masked, the model can easily fill the mask in a meaningful way. However, when each word is individually masked with a 15% chance it leads to a loss of context. After further experiments with increasing the mask to different values and finetuning, the model shows reasonable results at 70% masking (looks almost like a real tweet) while from higher than 75% masking, the output deteriorates. Hence, BERT represents a feasible technique for this purpose and a good basis for further research.

Mask %	Output
70	please read the most important thread by @ real _ @ @hfwmington explaining why # socialdistancing is in the name of the the # #securitying the the # coronavirus # covid19 outbreak.
75	# coronavirus # covid19 2 / 3 accelerate the #,,s, therapeutics and and and critical to # to to help all communities. contain misinformation minimize the spread spread.

Figure 4: Results in setting one

In the second experiment, two BERT models are combined into a GAN, see Figure 5. The BertForSequenceClassification model is used for the discriminator and the BertForMaskedLM model for the generator with an extra sigmoid layer added after the last linear layer of the BertForSequenceClassification model. GAN works by generating fake data with the generator which is sent to the discriminator that judges if real or not. Hence, the discriminator is trained with real texts labeled with one and fake text labeled with zero. To allow backpropagation of the loss from the discriminator to the generator, the first BertForSequenceClassification is replaced with a new layer with an extra check if the input is tokenized text, it is embedded as before, else if the input is already embedded, it passes as untouched. Additionally, the last decoder layer of BertForMaskedLM is replaced with an identity layer that passes the input to output without any transformation. Given the existing context, the results obtained are not outperforming the first experimental setting.

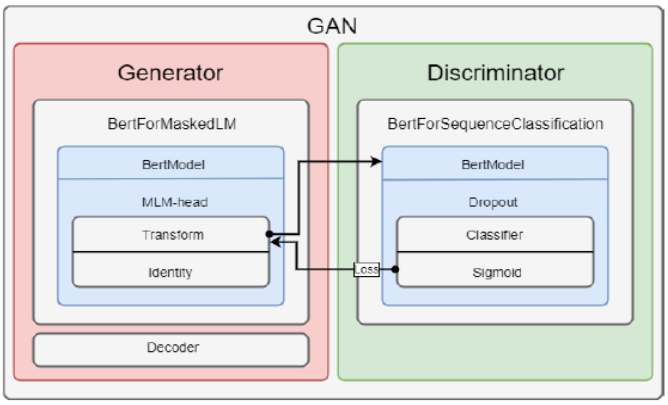


Figure 5: Model architecture setting two

In the third experiment depicted in Figure 6, the sigmoid layer from the first model is removed since BertForSequenceClassification already performs binary classification, the labels are provided in advance and a forward hook is added before the forward pass of the layer. From the experiments conducted, the generator is just as powerful as the discriminator, only slower. Moving this setting to a machine with less memory (6GB instead of 8 GB) shows a direct relation to the fact that although the generator is improving, it is slower than before.

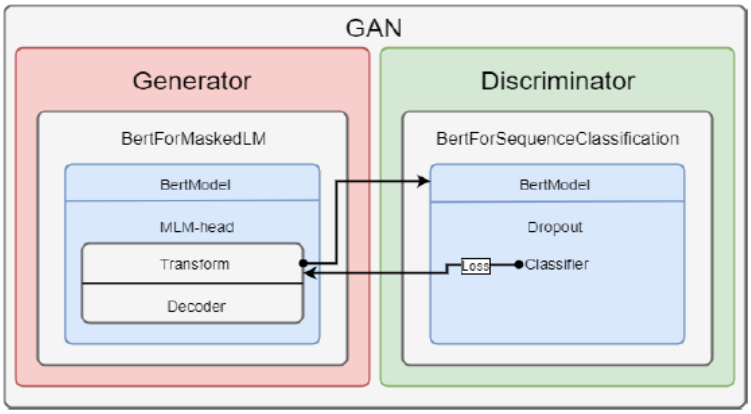


Figure 6: Model architecture setting three

Furthermore, focusing on the third setting, the hyperparameter tunning is applied. This mechanism implies improving the parameters that characterize the learning process, e.g., learning rate, batch size, using optimization techniques for obtaining optimal results. Accordingly, optimizers like SGD (Stochastic Gradient Descent which calculates the gradient of the loss function at random points) and Adam (updates network weights iteratively) are used. Then the model is trained on 10000 randomly selected tweets and validated on 1000 tweets. Hence, the generated tweets are analysed separately, and their metrics are compared with their original considering metrics like diversity for assessing text quality. Herein, while using only one optimizer, Rprop performs the worst, while the others well. Then, different combinations of optimization algorithms are tried which conducts to improved results as in Table 1. To assess which combination is the best, the RMSE (Root Mean Squared Error) of both the losses and the text quality metrics are calculated and shown in Table 2. Hence, the combination with the lowest RMSE for both text quality and the losses are Adam and SGD.

Table 1: Text quality of using combinations of optimizers

Optimizer Gen.	Optimizer Disc.	Word count	Diversity	Stop words	Punctuation count
Adam	AdaDelta	-2.58	-0.14	0.93	1.50
Adam	SGD	-2.36	-0.15	1.33	1.31
AdamW	AdaDelta	-2.23	-0.16	1.32	1.6
AdamW	SGD	-2.41	-0.16	1.49	1.27
RMSprop	AdaDelta	-1.38	-0.16	1.15	2.2
RMSprop	SGD	-1.52	-0.16	1.43	2.42

Table 2: Optimizer combination RMSE

Optimizer Gen.	Optimizer Disc.	Loss RMSE	Text Quality RMSE
Adam	AdaDelta	1.75	0.35
Adam	SGD	1.25	0.30
AdamW	AdaDelta	1.72	0.33
AdamW	SGD	1.25	0.32
RMSprop	AdaDelta	1.83	0.32
RMSprop	SGD	1.29	0.36

Moreover, when ignoring punctuation and stopwords so that they are never masked, the text quality significantly improves having the RSME value of 0.08 as in Table 3 with quite adequate generated tweets depicted in Table 4 and transformation of regular tweets into disinformation tweets shown in Table 5.

Table 3: Text quality after ignoring punctuation and stopwords

Word count	Diversity	Stop words	Punctuation count	Text Quality RMSE	Loss RMSE
-0.20	-0.02	0.03	0.28	0.08	0.90

Table 4: Original and generated tweets

Split	Original	Tweet
50/50	yes	when the #coronavirus first appeared, the natural public response was to demand that the government stop it. the next phase was to blame the government for failing to protect them. the third phase will be attacking the government for taking the steps it took to protect them. - gf
	no	when the # coronavirus outbreak emerged, the publicised reaction was to say that the chinese made it. the chinese reaction was to blame the government for trying to cure them. the real threat will be on the chinese for all the time it takes to cure them. - bsl
55/45	yes	tom hanks and his wife rita confirmed to have contracted #coronavirus in australia?? wow... not sure how reliable the sources are, but if it's true, damn.
	no	jalinhan and his team are supposed to have developed # coronavirus in canada??? ... not sure how real the rumors are, but if it's true, maybe.
60/40	yes	this should be the real worry surrounding the #coronavirus. the economic toll (and the unrest that comes with it) will be far worse than the death toll.
	no	this should be the cover story about the # coronavirus. the virus (and the people that worked with it) will be even worse than the lab lab.

Table 5: Transformation of regular tweets into disinformation tweets

yes	#healthminister: steps taken to prevent spread of #coronavirus in country
no	# coronavirus lab : samples sent to the laboratory of # coronavirus in canada
yes	how can we preserve the valuable networking experience from events as they move online? check out our latest blog post to learn more.
no	how can we keep the latest viruss from china as they come in? check out our own blogging to read more.

Conclusively, both the combination between GAN and BERT as well as BERT itself showed good results in disinformation data generation (hypothesis), and two facts should be considered: (i) with additional relevant data, the solution proposed could be further developed, and (ii) for reducing model's complexity, GAN could be eliminated from the architecture proposed.

6. Conclusions

The first step for building deterrence against social media manipulation mechanisms is security awareness tackling the existence and source plus the impact of akin incidents through corresponding solutions by protecting users' behaviour through cognitive or affective resistance and mitigating mechanisms like labelling (Bastick, 2022; Maathuis, Pieters & van den Berg, 2018b). Whilst such solutions often take defender's stance and are built on a limited set of publicly available social manipulation datasets, attackers improve through adaptivity and autonomy their skills and mechanisms for increasing the spread and intensity of their impact by either reaching quickly specific target audience or scaling it up for hitting a broader range of users when lacking specific target audience. This implies that building corresponding intelligent solutions for tackling social manipulation requires massive publicly available datasets and a multi-angle perspective (Kertysova, 2018; Maathuis, Pieters & van den Berg, 2021). Thus, as these represent major limitations in the existing body of knowledge, this research proposes a deep learning-based social manipulation security awareness approach by building an intelligent model based on GAN and BERT techniques as a disinformation dataset and corresponding generation solution that serves the purpose of strengthening ongoing research and development and reaching multi-domain audiences.

This research continues by combining this solution with a detection approach for building a comprehensive solution for strengthening the space of existing technical solutions dedicated to multi-domain audiences and which implicitly provides the basis for defining values, principles, and requirements that could be used when designing comprehensive social media security awareness solutions, policies, and strategies.

References

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004.

Auti, T., Sarkar, R., Stearns, B., Ojha, A. K., Paul, A., Comerford, M., ... & McCrae, J. P. (2022). Towards Classification of Legal Pharmaceutical Text using GAN-BERT. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, pp. 52-57.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in human behavior*, *116*, 106633.

Chen, L., Chen, J., & Xia, C. (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, *64*, 103060.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65.

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security. 17(1),* 26-35.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Delvin, J. & Chang, M. W. (2018). Open source BERT: state-of-the art pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

EU Commission (2018). Communication-Tackling online disinformation: a European approach. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-online-disinformation-european-approach

EU Commission (2022a). Tackling online disinformation. https://digital-strategy.ec.europa.eu/en/policies/online-disinformation.

EU Commission (2022b). The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

Ki-Aries, D., & Faily, S. (2017). Persona-centred information security awareness. *computers & security*, *70*, 663-674.

Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., & Choi, D. (2021). FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics*, *64*, 101688.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1-4), 55-81.

Kula, S., Kozik, R., & Choraś, M. (2021). Implementation of the BERT-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*, 1-13.

Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, 282-291, IEEE.

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. (2017). Data-driven news generation for automated journalism. In *Proceedings of the 10th international conference on natural language generation*. 188-197.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2016). Cyber weapons: a profiling framework. In *2016 International Conference on Cyber Conflict (CyCon US)*, pp. 1-8, IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). Developing a cyber operations computational ontology. *Journal of Information Warfare*, *17*(3), 32-49.

Maathuis, C., Pieters, W., & Van den Berg, J. (2018b). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pp. 1-6, IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2021). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*, *17*(2), 352-374.

Maathuis, C. (2022a). On Explainable AI Solutions for Targeting in Cyber Military Operations. In *International Conference on Cyber Warfare and Security*. 17(1), 166-175.

Maathuis, C. (2022b). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security*. 21(1), 170-177.

Maathuis, C., & Chockalingam, S. (2022b). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security.* 21(1).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

NATO (2021a). Countering disinformation: improving the Alliance's digital resilience. https://www.nato.int/docu/review/articles/2021/08/12/countering-disinformation-improving-the-alliances-digital-resilience/index.html

NATO (2021b). How does NATO respond to disinformation? https://www.nato.int/cps/en/natohq/news_184036.htm

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuition*, 21-29, Springer, Cham.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Pérez-Escolar, M., Ordóñez-Olmedo, E., & Alcaide-Pulido, P. (2021). Fact-Checking Skills and Project-Based Learning about Infodemic and Disinformation. *Thinking Skills and Creativity*, *41*, 100887.

Rathenau (2022). AI and manipulation on social and digital media. https://www.rathenau.nl/en/digital-governance/ai-and-manipulation-social-and-digital-media

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, *291*(3), 906-917.

Shu, K. (2022). Combating disinformation on social media: A computational perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 100035.

Shu, K., Li, Y., Ding, K., & Liu, H. (2021). Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 35(15), 13825-13833.

Sonkiya, P., Bajpai, V., & Bansal, A. (2021). Stock price prediction using BERT and GAN. *arXiv preprint arXiv:2107.09055*.

Ta, H. T., Rahman, A. B. S., Najjar, L., & Gelbukh, A. (2022). GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media.

Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2017). Choosing a design science research methodology.

Wang, X., Guo, H., Hu, S., Chang, M. C., & Lyu, S. (2022). Gan-generated faces detection: A survey and new perspectives. *arXiv preprint arXiv:2202.07145*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu, G., Yang, Y., Wang, X., Zhen, H., He, G., Li, Z., ... & Shu, L. (2020). Adversarial active learning for the identification of medical concepts and annotation inconsistency. *Journal of Biomedical Informatics*, *108*, 103481.

Social Media Manipulation Deep Learning based Disinformation Detection

Clara Maathuis[1], Rik Godschalk[2]
[1]Open University of the Netherlands, Heerlen, Netherlands.
[2]Independent Researcher, Netherlands.

clara.maathuis@ou.nl
ahgodschalk@gmail.com

Abstract: The rapid grow and use of different social platforms enhanced communication between different entities and their audiences plus the transformation through digitalization of existing, e.g., ideas and businesses, or the creation of new ones fully existing or depending on this digital environment. Nevertheless, next to these promising aspects, social media is a vulnerable digital environment where a diverse plethora of cyber incidents are planned and executed engaging a diverse range of targets. Among these, social media manipulation through threats like disinformation and misinformation produce a broad span of effects that cross digital borders into the human realm by influencing and altering human believes, behaviour, and attitudes towards specific ideas, institutions, or people. To tackle these issues, existing academic, social platforms, dedicated organizations, and institutions efforts exist for building specific advanced and intelligent solutions for detecting and preventing them. Regardless, these efforts embed defender's perspective and are focused locally, at target level, without being designed to fit a broader agenda of producing and/or strengthening social media security awareness. On this behalf, this research proposes a deep learning-based disinformation detection solution for facilitating and/or enhancing social media security awareness in respect to offender's perspective. To achieve this objective, a Data Science approach is taken based on the Design Science Research methodology, and the results obtained are discussed with a keen on further field developments regarding intelligent, transparent, and responsible solutions countering social manipulation through realistic participation and contribution of different stakeholders from different disciplines.

Keywords: information operations, cyber operations, social manipulation, disinformation, misinformation, security awareness, machine learning, deep learning.

1. Introduction

"I do respect people's faith, but I don't respect their manipulation of that faith in order to create fear and control." (Javier Bardem)

The efforts that agents like social platforms and dedicated outlets, institutions, researchers, and practitioners invest for dealing with social media manipulation mechanisms imply firstly understanding what they mean, why and how are built, and what is the scale and intensity of their force projection at individual and collective levels (Bastick, 2021; Fard & Maathuis, 2021). For instance, wars like the ones in Ukraine and Syria reflect how social media is used for communication purposes next to, e.g., intelligence information gathering, influencing public perception through techniques like propaganda, attacking the opposition or launching aggressive online campaigns, driving division and polarization, and building psychological deterrence through trolling, harassment, and personal attacks (Chen, Chen & Xia, 2022). Another example is the ongoing Covid-19 pandemic which through its digital presence and impact scale received its corresponding digital name Infodemic seeing the superabundance of information regarding concepts and events surrounding it, e.g., nature of the virus, its effects, vaccination types and campaigns, unknown and unforeseen effects of vaccines, and relation between the virus, vaccination, and general effects affecting people (WHO, 2020; Pérez-Escolar, 2021; Europol, 2021; Ivendi, 2022; Chockalingam & Maathuis, 2022). Hence, while defending mechanisms should be properly designed, developed, and deployed, they should consider a security awareness goal on existing and possible future threats plus the necessity of protecting (socio-ethical) values of humans and systems (EU Commission, 2022b; Maathuis & Chockalingam, 2022c).

Accordingly, this research aims to build a social manipulation detection solution that takes offender's perspective through a deep learning model for building or strengthening social media manipulation security awareness through a detection approach (Kaliyar, Goswami & Narang (2021), i.e., based on both context and content. Hence, a multidisciplinary perspective is adopted by integrating relevant studies from cyber/information security and operations, social media, and deep learning following a Design Science Research methodological approach (Peffers, 2007; Peffers, 2018) through the Data Science framework having the following contributions:

- To research and practitioner communities that could adapt/change their approach to thinking through the eyes of the attacker/offender when building social manipulation countering initiatives and solutions.
- To security, social media, and AI experts as the solution proposed is implemented based on deep learning techniques by encouraging them to consider creative and advanced approaches when dealing with social manipulation, and if applicable extend the approach taken in this research, or further release useful social manipulation datasets that could be used in upcoming studies.
- To decision makers involved in designing, developing, and deploying strategies and solutions that produce/enhance social media awareness on manipulation mechanisms and incidents based on intelligent techniques that should be further understood by a multi-domain audience and approach.

The reminder of this article is structured as follows. Section 2 discusses the background of this research together with relevant studies. Section 3 describes the research approach taken. Section 4 presents the deep learning techniques considered for building the model proposed. Section 5 discusses model implementation reflecting on choices and the results obtained. Section 6 discusses the findings and future research lines.

2. Social Manipulation Detection Research

The UN stresses that "information pollution is a hard-to-fix problem, for which there is no easy solution" and that disinformation is an "existential risk to humanity" as it damages trust in democratic societies, governments, institutions, and processes further calling to division and conflict (UN, 2021a). Specifically, the Common Agenda (UN, 2021b) contains 12 commitments from where the third (Promote peace and prevent conflicts), fourth (Abide by international law and ensure justice), sixth (Build trust), and seventh (Improve digital cooperation) are the most relevant measures when dealing with disinformation (UN, 2021c). The EU Commission highlights how important is to protect democratic values by tackling social manipulation methods like disinformation which is misleading or false information spread through social platforms for deceiving or protecting specific economic or political gain (EU Commission, 2022a). The institution proposed a set of initiatives on combating disinformation like the European Democracy Action Plan and the Strengthened Code of Practice on Disinformation which argues for, e.g., strengthening political transparency and empowering users and researchers (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b). Moreover, through the risk assessment on Covid-19 disinformation conducted by Europol, resulted that disinformation threats are represented by several agents, i.e., individuals like criminals that want to get profit, states and state-backed actors that hope to achieve specific geopolitical aims, and opportunists who want to discredit official (re)sources.

Reisach (2021) argues that social platforms allow carrying out diverse businesses and political discourses, but also play a role in the creation and execution of social manipulation campaigns while integrating AI techniques to perform different functions. Furthermore, the author argues that society, social media platforms, and their users share their responsibility for users' (in)actions. To support this, the author suggests developing effective education solutions that are geared toward helping users become more responsible and tech-savvy. These include providing educational support to vulnerable populations, developing an ethical mindset, and promoting tech and media literacy. Chen, Chen, and Xia (2022) investigate user behavior and regulation processes in social networks and argue that social media has become an environment for the use of cyber weapons (Maathuis, Pieters & van den Berg, 2016), thus a new battlefield environment (Maathuis, Pieters & van den Berg, 2018c). Herein, power is augmented by goals that include analyzing relationships between people and actively intervening with deceptive techniques. Additionally, the authors are concerned since social media is used for warfare purposes or just as infosphere of cyber weapons, e.g., bots, botnets, trolls, and cyborgs with their functioning modules that include behaviour representation, user recognition, and discovery of homogeneous communities. Moreover, Shu (2022) performs a survey on existing recent detection advancements while acknowledging still one of the biggest issues in this field: the data challenge, i.e., the limited or lack of publicly open datasets which could be helpful for developing solutions for understanding existing manipulation topics, methods, and tactics utilized. Furthermore, the author acknowledges the role and increase in spread of deep learning-based solutions, but argues that when building and deploying them, they should be trustworthy, i.e., based on human knowledge, take into consideration aspects like noise or incomplete human feedback for improving representation learning, and broadly interpret the results obtained.

Critical to understanding and dealing with such social manipulation mechanisms is the existence and availability of relevant datasets as this represents the basis for building tackling/countering solutions. Patwa et al. (2021) built a Twitter fake news dataset with real and fake labels and a set of corresponding baseline ML (Machine

Learning) models. Furthermore, Kim et al. (2021) proposed the FibVID (Fake News Distribution Database for Covid-19) containing tweets with both Covid-19 and non-Covid-19 fake news content to support further analysis of the phenomenon and its spreading mechanisms. Sharma, D. K., & Garg et al (2021) proposed the IFND (Indian fake news dataset) dataset which has text and images for fake news identification based on fact-checking events from India between 2013-2021.

On detecting disinformation, Nasir, Khan & Varlamis (2021) propose a fake news solution based on a hybrid deep learning approach embedding CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) techniques obtaining good results in comparison with other recent models. Moreover, Kaliyar, Goswami & Narang (2021) argue for capturing semantic and long-distance dependencies in sentences and build a hybrid deep learning model combining BERT (Bidirectional Encoder Representations from Transformers) with CNN techniques in a model named FakeBERT for capturing fake news relevant information for improving the classification performance. The authors stress that such an architecture helps dealing with ambiguity which represents one of the main challenges in this field. Seeing the high accuracy obtained (98.90%) and the approach followed, this research considers a similar combination. As sentiment analysis proved useful when detecting social manipulation mechanisms, Drus & Khalid (2019) execute a systematic review on different methods and conclude that both lexicon-based or ML-based techniques (baseline techniques like Naïve Bayes and SVM (Support Vector Machine), and deep learning) plus a combination thereof are used for social media analytics tasks ranging from consumer trends, disaster impact assessment, feelings and perceptions surrounding the ISIS group, to governmental elections. Furthermore, focusing on the common points of the applicability of diverse deep learning techniques, e.g., CNN and LSTM for multilingual sentiment analysis, Torales, Salas & Herrera (2021) militate for focus changing from single language to multilingual analysis, and the upcoming trend in building transformer-based architectures like the ones we implement.

The term Infodemic is defined by WHO (World Health Organization) as “too much information including false or misleading information in digital and physical environments during a disease outbreak” (WHO, 2020). Ivendi et al. (2022) conduct a sentiment analysis-based research considering data fusion from news broadcasting, health, and governments websites on different issues and events surrounding the Covid-19 pandemic. The authors use features like positive, negative, or neutral categorizations, text language and character count. For implementation, a deep learning approach is taken based on RNN and two-word clouds for real and real news with words like vaccine, people, and health, and fake news with China, pandemic, and Wuhan exist. Based on the results obtained the anger, fear, surprise, and negative emotions have higher frequency values in the fake news content when compared to the real news content. Moreover, to investigate sentiments in social media discussions on Covid-19 vaccination, a sentiment analysis and LDA (Latent Dirichlet Allocation) topic analysis is conducted by (Melton et al., 2021) based on Reddit data from December 2020 to May 2021. There the results show that in general a more positive sentiment is present in vaccine-related discussions over time although as the topic modelling shows, hesitation on vaccination implementation, digital interventions, and building new policies exist.

From the literature review conducted, this research acknowledges the limited amount of open publicly available datasets plus the risks associated with their availability, and the incipient state of the existing body of knowledge on advanced/ hybrid deep learning approaches for tackling social manipulation.

3. Research Methodology

As this research intends to build an intelligent model that detects social media manipulation mechanisms like disinformation for further contributing to building/enhancing social media security awareness of various communities, the following research questions are considered:

- How to build a deep learning model for detecting disinformation considering producing social media security awareness?
- How to evaluate the proposed model?

To reach this goal, a Design Science Research approach (Peffers, 2007; Peffers, 2018) is followed based on the Data Science framework through multidisciplinary research (Venable, Pries-Heje & Baskerville, 2017; Maathuis, Pieters & van den Berg, 2018b). Hence, a deep learning-based model is developed, evaluated, and proposed considering the following research activities:

Problem definition and solution aim: while the existing strategies, policies, and solutions consider, define, and implement different mechanisms for preventing and detecting social manipulation, they usually (i) take

defender's perspective which in this field proved weak as the offender has advantage in attack, and (ii) are still reliant on limited data and/or human experts which could be incomplete or might imply unexpected (propagated)errors/biases. Furthermore, seeing the large-scale impact and discourses that the ongoing war in Ukraine and Covid-19 pandemic continue to produce, there is an increasing need for building advanced and transparent intelligent solutions (Maathuis, Pieters & van den Berg, 2021) for proactively detecting social manipulation through offensive lenses. Thus, this research adopts offender's perspective proposing a deep learning-based social manipulation detection solution. Accordingly, literature review is conducted in scientific databases like IEEE Digital Library, ACM Digital Library, Scopus, and Google Scholar using combinations of keywords like 'social media', 'manipulation', 'disinformation', 'detect', and 'deep learning'.

Solution development: to develop the model, Covid-19 pandemic tweets are collected focusing on the first two waves, from 01.01.2020 to 21.03.2020 (first wave) and from 15.10.2020 to 31.12.2020 (second wave). Furthermore, the processes of data collection, processing, hashtag analysis, topic extraction, development using CNN and BERT, and the evaluation mechanism are implemented in Python and are further addressed. Specifically, the number of times used per hashtag is depicted in Figure 1:

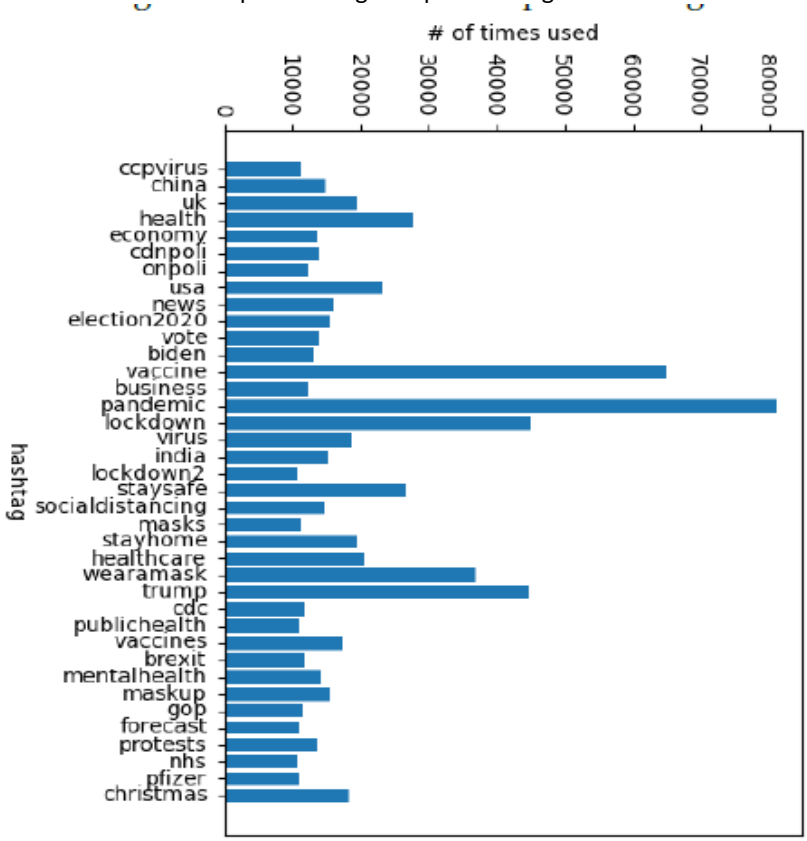


Figure 1: Number of times used per hashtag

Solution evaluation: the model is assessed evaluation metrics like accuracy and is improved through optimization while the results obtained are discussed.

Communication: the results obtained are discussed here and integrated in ongoing research in this domain while addressing the need for (i) realising publicly available social manipulation datasets for research and practitioner communities, and (ii) merging human with technological assessment in modelling and simulation settings for proactively countering social manipulation campaigns/operations, mechanisms, and incidents.

4. CNN and BERT Modelling Techniques

Due to the increase in innovation, computation, and (multi-source) data availability, the developments and solutions proposed based on AI techniques that match and even go further beyond human cognition and decision-making through different ML approaches, either have completely changed some societal domains or are in the process of doing that (Janiesch, Zschech & Heinrich, 2021; Maathuis, 2022b). Currently, at the core of development in the field of ML is building intelligent (complex) neural architectures, i.e., deep learning, which

while is being perceived as a novelty, "is around for decades" (Ng, 2017). A deep learning architecture contains three or more layers, meaning input, hidden, and output, that learns from a series of mathematical representations and operations from (large amounts of) data (Goodfellow, Bengio & Courville, 2016). Since this research uses specific neural architectures, i.e., BERT – Bidirectional Encoder Representations from Transformers and CNN – Convolutional Neural Networks, they are further introduced for background purposes.

CNN is among the most used deep learning techniques, was introduced by LeCun et al. (1989), and has a hierarchical structure implying extracting features from data through convolutional operations, see Figure 2. A CNN contains the following layers: input, convolution, pooling, fully connected, and output. The operational flow is as follows: convolutional layers learn features from the input data depending on the kernel used and pass results to pooling layers where data is condensed to a lower dimensional data without losing much information and further provided to fully connected layers which are normal fully connected neural networks. Shortly, the convolutional and pooling layers are used for extracting feature information from the dataset before the data is used to train the network for a task (O'Shea, K., & Nash, 2015; Gu et a., 2018; Li et al., 2021). Additionally, multiple pairs of convolutional and pooling layers could exist and learning is done through backpropagation.

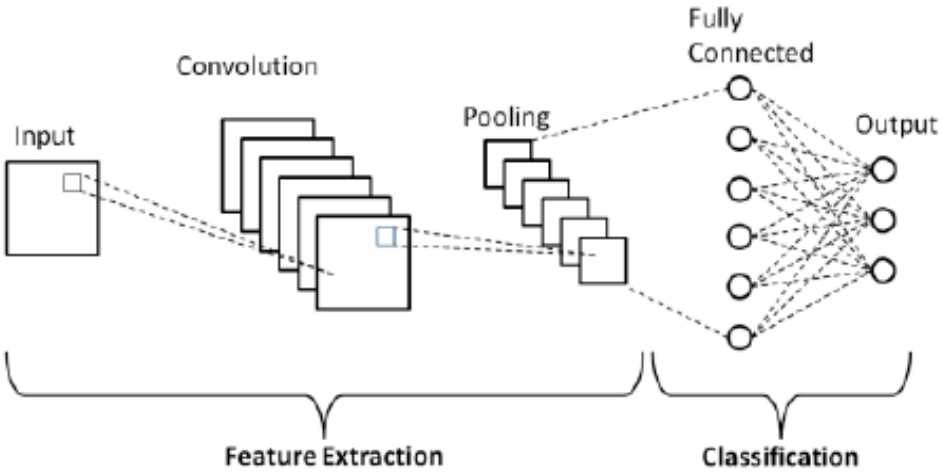


Figure 2: CNN architecture from (Phung & Rhee, 2019)

BERT is a transformer-based deep learning technique proposed by Delvin et al. (2018) and embeds twelve layers where through each layer more contextual information is acquired. Through its nature, BERT is the first left-to-right and right-to-left language architecture (Figure 3), i.e., deeply bidirectional. The input layer is the tokenized data, and the output layer is the masked language model which makes it a good choice when building text contextual embedding and text classification solutions (Delvin & Chang, 2018; Qiu et al. 2020).

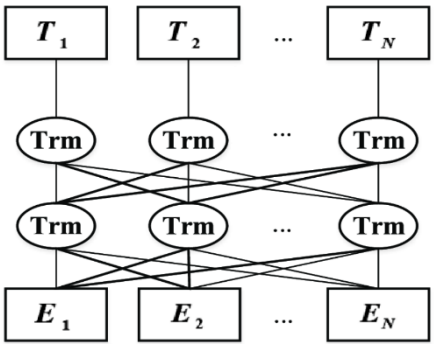


Figure 3. BERT model from (Lu et al. 2020)

5. Model Implementation and Evaluation

Previously, solutions that mix BERT with CNN show good results for tasks like depression detection (Rodrigues Makiuchi et al., 2019), message credibility assessment for fake news detection (Verna et al., 2022), and offensive speech identification in social media (Safaya, Abdullatif & Yuret, 2020). In this research BERT is combined with CNN with an additional feature from sentiment analysis that captures emotions within text. Moreover, for

implementation the Huggingface (Wolf et al., 2019) and VADER (Hutto & Gilbert, 2014) libraries are used in Python for BERT and generating polarity values for the twitter messages that are used as sentiment values, respectively. The evaluation is made using Covid-19 tweets.

The techniques are combined based on the architecture depicted in Figure 3 where the data flow and reasoning go from the input data(tweets) to output which represents classifying the input data as disinformation or not. Herein, a CNN model is added on top of the BERT layers with the architecture of the CNN model laid out in with ReLu (rectified linear units) as activation function (Chen, 2015) where the following changes are applied: (i) the number and size of convolutional layers is made a hyperparameter for allowing additional testing of size combinations to see which one is suitable for this goal, (ii) the softmax function is not used as the model only predicts two classes, and (iii) providing additional data into the CNN during the training phase as a way to add sentiment values as feature. This results in a vector with an additional dimension, step which could be skipped if desired which implies that no extra sentiment information will be added and later can be assessed if its addition improves the results obtained.

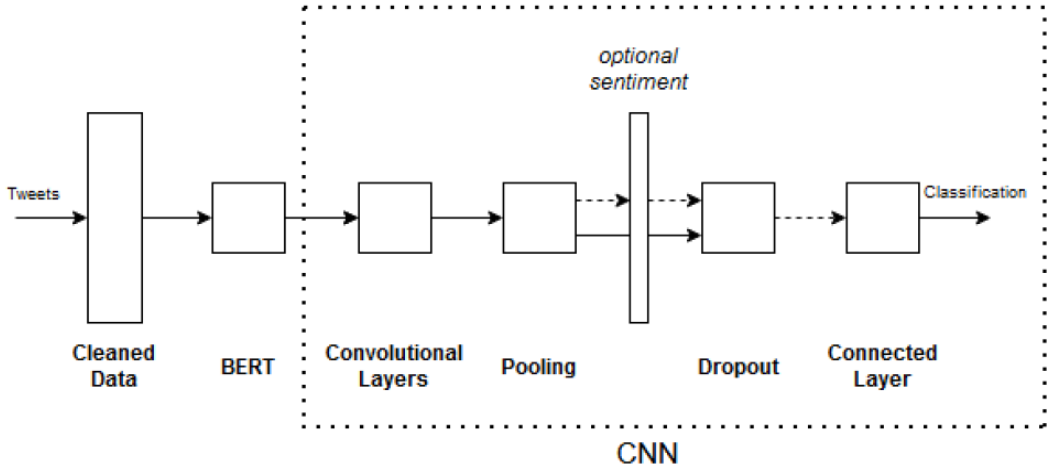


Figure 3`; Model architecture

Moreover, the hyperparameters of the model are: dropout (probability that a feature is dropped by the input layer), epochs (number of epochs to train the model), learning rate (rate at which the internal weights get adjusted each training step), sentiment (if the model should use the sentiment feature), convolution windows (list of sizes for the different convolutional windows), and bert layers. Accordingly, if the hyperparameter sentiment is set to true then extra sentiment values are unpacked from the batch and added to the call to the CNN network; otherwise, if the value is false then only the hidden layers from the BERT network are passed. Shortly, the model is trained during a given number of epochs that each consist out of smaller steps where the amount of these depends on the batch size and the size of the training set. During each step data from a batch is passed through both the BERT and CNN network which gives a class prediction of a sample. This sample is then compared to the actual class to calculate the loss which is in turn used to tune the weights in both networks. Once all epochs are completed the model is fully trained. Since there are many combinations of hyparameters that could be used for tunning, a random selection is adopted to reduce training and evaluation time. Hence, the results are in Table 1 and show that the model already reaches a high score after completing the second epoch and continues training for two more where it improves. Even though the model has a high accuracy, this does not necessarily mean that is the best option since it could overfit to training data and could not perform well on test data.

Table 1: Results overview

Epoch #	Training Loss	Training F1	Training Time	Validation Accuracy	Validation F1
1	0.189	0.938	0:00:30	0.934	0.920
2	0.033	0.988	0:00:34	0.989	0.993
3	0.01	1.00	0:00:33	0.989	0.993
4	0.01	0.998	0:00:36	0.89	0.993

For testing the model, two perspectives are considered: the first with sentiment values as extra feature and the second without. Accordingly, in Table 2 and Table 3 the results for three sets for both perspectives are presented. Hence, both perform at a similar level, but it appears that using sentiment values might improve the

generalization towards unseen data for the model reflecting the importance of human-based contributions.

Table 2: Test results with sentiment feature

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.165	0.958	0.957	0.958	0.951
Test #1	0.108	0.979	0.963	0.979	0.970
Validation #2	0.076	0.969	0.968	0.969	0.964
Test #2	0.099	0.969	0.977	0.969	0.965
Validation #3	0.141	0.967	0.975	0.969	0.967
Test #3	0.088	0.979	0.971	0.979	0.973

Table 3: Test results without sentiment feature

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.050	0.990	0.981	0.990	0.985
Test #1	0.080	0.979	0.981	0.979	0.971
Validation #2	0.036	0.979	0.990	0.979	0.998
Test #2	0.080	0.979	0.963	0.979	0.970
Validation #3	0.07	0.989	0.980	0.990	0.984
Test #3	0.154	0.958	0.972	0.958	0.962

To further test the quality of the model, a randomly selected sample of 10000 tweets is taken with 9874 non-empty tweets new to the model. Out of them, 4 were labelled as disinformation as in Table 4: while the first and fourth are clearly disinformation, the second and third tweet could be more nuanced seeing the training set. After a manual check on random tweets (thus not all due to its size), we conclude that do not contain disinformation. Of course, further updates can be considered when more data are available, and new forms of disinformation forms are identified.

Table 4: Disinformation tweets

Tweet
Leak originated from a bio weapons lab in Wuhan which I believe is in China.
The truth behind delayed and broken #coronavirus CDC testkits tied to contamination due to poor mishandling in Atlanta lab, says federal scientist.
Hackers posing as exec from #China based Haier Biomedical target #COVID19 vaccine cold-chain, hoping to disrupt distribution. Potentially the work of a government backed entity, given targeting methods. Report from IBM
Corona Virus is a biological distraction created by the USA To disguise the financial crash that has been triggered by the FED to burst to equity bubble. This is a Black Swan event hidden from the public by the Elite.

6. Conclusions

As long as the humankind exists, conflict and manipulation are present, thus building a broad range of skills transferred/transformed to strategies, tactics, and solutions for tackling threats like manipulation inside or through social media platforms, and facilitating the development of intelligent, autonomous, and adaptive threats for producing new or enhancing existing social manipulation mechanisms. Accordingly, for developing next generation or enhancing existing solutions that tackle social manipulation mechanisms like disinformation, a high number of diverse datasets and a multi-faceted approach are necessary (Tucker et al., 2018; Maathuis, Pieters & van den Berg, 2018) while keeping humans in the loop. To tackle these issues, this research proposes a disinformation detection solution based on combining deep learning techniques, BERT and CNN with an additional feature that captures the sentiment values taking the offender's perspective and embedding lessons learned from previous social media manipulation campaigns. To build this solution, a Design Science Research is conducted through a Data Science approach using a built-in dataset of Covid-19 tweets.

As the solution proposed provides valuable results to the existing body of knowledge and considering existing practitioner approaches, this research advances by merging this solution with a generative approach for building a social media security awareness system that incorporates both defender and offender perspectives, and from there proposing design principles and requirements useful when building systematic corresponding strategies, policies, mechanisms, and solutions.

References

Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, *107*, 107373.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in human behavior*, *116*, 106633.

Chen, Y. (2015). *Convolutional neural network for sentence classification* (Master's thesis, University of Waterloo).

Chen, L., Chen, J., & Xia, C. (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, *64*, 103060.

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security.* 17(1), 26-35.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Delvin, J. & Chang, M. W. (2018). Open source BERT: state-of-the art pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707-714.

EU Commission (2018). Communication-Tackling online disinformation: a European approach. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-online-disinformation-european-approach

EU Commission (2022a). Tackling online disinformation. https://digital-strategy.ec.europa.eu/en/policies/online-disinformation.

EU Commission (2022b). The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

Europol (2021). Covid-19 fake news. https://www.europol.europa.eu/covid-19/covid-19-fake-news

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Ferrara, E. (2020). # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354-377.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media.* 8(1), 216-225.

Iwendi, C., Mohan, S., Ibeke, E., Ahmadian, A., & Ciano, T. (2022). Covid-19 fake news sentiment analysis. *Computers and electrical engineering*, *101*, 107967.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.

Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., & Choi, D. (2021). FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics*, *64*, 101688.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

Lu, Q., Zhu, Z., Xu, F., Zhang, D., Wu, W., & Guo, Q. (2020). Bi-gru sentiment classification for chinese based on grammar rules and bert. *International Journal of Computational Intelligence Systems*, *13*(1), 538-548.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2018b). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security*, pp. 278-288.

Maathuis, C., Pieters, W., & Van den Berg, J. (2018c). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, 1-6, IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2021). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*, *17*(2), 352-374.

Maathuis, C. (2022a). On Explainable AI Solutions for Targeting in Cyber Military Operations. In *International Conference on Cyber Warfare and Security.* 17(1), 166-175.

Maathuis, C. (2022b). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security.* 21(1), 170-177.

Maathuis, C., & Chockalingam, S. (2022c). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security.* 21(1).

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505-1512.

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.

Ng, A. (2017). Machine learning yearning. *URL: http://www. mlyearning. org/(96)*, *139*.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuition,* 21-29, Springer, Cham.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Rodrigues Makiuchi, M., Warnita, T., Uto, K., & Shinoda, K. (2019, October). Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (pp. 55-63).Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Pérez-Escolar, M., Ordóñez-Olmedo, E., & Alcaide-Pulido, P. (2021). Fact-Checking Skills and Project-Based Learning about Infodemic and Disinformation. *Thinking Skills and Creativity*, *41*, 100887.

Phung, V. H., & Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, *9*(21), 4500.

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, *291*(3), 906-917.

Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation,* 2054-2059.

Sharma, D. K., & Garg, S. (2021). IFND: a benchmark dataset for fake news detection. *Complex & Intelligent Systems*, 1-21.

Shu, K. (2022). Combating disinformation on social media: A computational perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 100035.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature.*

UN (2021a). Countering disinformation and promoting data transparency. https://www.undp.org/speeches/countering-disinformation-and-promoting-data-transparency?utm_source=EN&utm_medium=GSR&utm_content=US_UNDP_PaidSearch_Brand_English&utm_campaign=CENTRAL&c_src=CENTRAL&c_src2=GSR&gclid=EAIaIQobChMI7t2Di8KC-gIVRJBoCR38vQajEAAYAyAAEgLBJfD_BwE

UN (2021b). Common Agenda. https://www.un.org/en/common-agenda

UN (2021c). UN Common Agenda. https://www.un.org/en/content/common-agenda-report/

Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2017). Choosing a design science research methodology.

Verma, P. K., Agrawal, P., Madaan, V., & Prodan, R. (2022). MCred: multi-modal message credibility for fake news detection using BERT and CNN. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.

WHO (2020). Infodemic. World Health Organization. https://www.who.int/health-topics/infodemic#tab=tab_1

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Improvements on Hiding x86-64 Instructions by Interleaving

William Mahoney[1], J. Todd McDonald[2], George Grispos[1] and Sayonnha Mandal[1]
[1] School of Interdisciplinary Informatics, University of Nebraska at Omaha, Omaha, Nebraska, USA
[2] Computer Science Department, University of South Alabama, Mobile, Alabama, USA

wmahoney@unomaha.edu
jtmcdonald@southalabama.edu
ggrispos@unomaha.edu
smandal@unomaha.edu

Abstract: This paper presents the results of a new method for interleaving CPU instructions in x86-64 machine code, such that one can hide executable code within other valid instructions. The aim is to make it more difficult to reverse-engineer software at a machine code level – to obfuscate instructions. A result is a hidden execution path within a visible main instruction path. While previous methods for this instruction obfuscation exist, we present a new method which builds upon past work, and which allows a greater flexibility in the selection of main instruction path instructions. The result of this new approach is to provide a methodology for instruction concealment which is free of restrictions present in prior work.

Keywords: Instruction selection, instruction set architecture, obfuscation, reverse engineering

1. Introduction

The process of reverse engineering software makes the software susceptible to various types of attacks if vulnerabilities can be discovered, intellectual property theft if key algorithms can be discerned, and violating service agreements if portions of licensing software can be disabled. In some cases, therefore, software authors may be motivated to make the reverse engineering process more difficult. Looking at things from the opposite perspective, malware authors also wish to make their malware difficult to reverse engineer. Regardless of the motive, a methodology which has been explored in the past is a form of "executable steganography"; one can embed machine instructions which can be executed but which are not obvious when reverse engineering the program.

Steganography involving instructions hidden in other data is simply a specialization of steganography in general. Numerous techniques already exist and are available for embedding digital data into image files, for example (Morkel 2005) and (Hamid 2012). If, using steganography, a picture can hide data that includes executable software, then it is reasonable to think that one can do the opposite and hide an image inside of executable code. And for that matter, one should be able to conceal executable code inside of executable code.

We are not the first researchers to propose executable programs as cover objects for steganographic purposes. Other researchers have explored executable steganography for hiding not only other programs but other data in general. One previous approach to executable steganography in particular will be described and reviewed in this paper, as this research uses the method as a starting point, presenting potential modifications and thoughts concerning the method. Specifically, we will refer at length to a technique described by Christopher Jämthagen, Patrik Lantz, and Martin Hell in 2013 (Jamthagen 2013). These authors present an algorithm involving what they refer to as a Main Execution Path (MEP), as well as a Hidden Execution Path (HEP). The goal of their method is to hide HEP within MEP according to certain constraints. In our research we will aim to provide improvements to the method, which we will refer to as "JLH" for Jämthagen, Lantz, and Hall.

The paper here initially describes, in section two, the common methods used to recover instructions via reverse engineering. These methods are typically static analysis (not executing the program), although the analysis may need to happen at runtime if code is dynamically unpacked or unencrypted at runtime.

Since we are specifically using JLH as our starting point, we forgo much of the traditional "prior work" section and instead present the background of their method in section three. Once an understanding of JLH is complete in section three we turn to and describe a related approach involving the scattering of instructions within an array of bytes, and the addition of MEP instructions using a depth-first approach. This approach is in section four. Section five provides some example results and thoughts, and conclusions and future work are in section six.

2. Reverse Assembling Binary Code

The problem of reverse engineering binary code comes down to reverse assembling the object code and instructions back into something that is human readable. In some cases this is easier than in other cases. For example, if the binary executable is for an ARM or other typical Reduced Instruction Set Computer (RISC), the process is helped by the fact that the instructions are typically a fixed length and "aligned" on certain addresses. In the case of Complex Instruction Set Computers (CISC) and the x86 family, the reverse assembly process is quite complex and cumbersome. Even the number of, and types of these instructions can be a mystery (Mahoney 2021).

Static analysis, looking at the instructions without executing the program, develops or discovers properties that hold true for all executions of a program. De-compilation (or reverse compilation), disassembly, control flow discovery, and data flow discovery are some examples of static analysis. Contrasting this is dynamic analysis, where the program is executed and data is collected. This leaves potential undiscovered pathways within the code; these paths will not be analyzed if they were not executed on a particular run of the program. Popular tools for static analysis include Ghidra (2022), objdump (LinuxDevCenter 2022), and udcli (Udis86 2022). The static analysis tool IdaPro (HexRays 2022), performs static analysis but also integrates with various debuggers, making this tool partially static and partially dynamic. Dynamic analysis tools include Binary Ninja (2022), Hopper (2022), OllyDbg (2022), Relyze (2022), x64dbg (2022), and more. Online web sites are also available for disassembly, including simple interfaces provided by – for example – Defuse (2022). All of these tools use several approaches to determine how the code operates.

Some are static analysis tools only, while some include integrated debuggers. Udcli and objdump programs are simple command-line static disassembly tools, while others such as OllyDbg and x64dbg are more traditional debugging platforms which are used mainly for dynamic analysis.

Whether the reverse assembly process is dynamic or static, there are two general approaches to reverse engineering the instruction stream: a linear scan method or a tree-based method (Linn 2003). Scanning starts at a program entry point and assumes that the instructions are contiguous. A scan simply converts the entire instruction area back to assembly language, translating one instruction at a time until there are no more instructions. A tree-based approach also starts at the entry point of the program, but then uses intelligence to queue the destinations of branch/jump instructions, starting the reverse assembly process at each destination. Simple reversing tools such as the Linux objdump utility perform a linear scan only, while advanced tools such as IdaPro use a combination of linear and tree-based scanning. If the target of a jump instruction is known, the reversal process is relatively straightforward, but branches to addresses that are calculated at runtime will not be known with static analysis and as a result, some code may be missed. Additionally, the x86 architecture includes several different categories of jump instructions (AMD 2022), (Intel 2006), including unconditional or conditional branches, a relative or absolute address as a destination, and may be indirect (a register holding the address). It is thus not possible to determine the target of every jump using only static analysis.

In the work described here we are primarily concerned with placing the hidden execution path (HEP) within the main execution path (MEP) in such a way that traditional reversing tools such as IdaPro (Pearce 2008) are confused into thinking that the MEP is the correct code. This is the aim of the approach described in JLH, but their approach has a very restricted view for selecting the instructions used for the MEP. Nonetheless it is necessary to next describe their approach, prior to suggesting improvements or changes.

3. The Jämthagen, Lantz, and Hall Method

3.1 Background: Anti-Disassembly

A well-known and understood technique for thwarting the reverse-assembly process (at least on x86) is the insertion of "junk" bytes to throw off the instruction decoding process. We have reported on this in the past and related as to which additional bytes of junk cause the most damage in terms of misrepresented instructions (Shinn, Mahoney 2011). An example is shown in Figure 1 below from Mahoney (2015), where the addition of 0x05 has caused the reverse-assembly process to mis several instructions.

If the inserted extraneous bytes ("junk" bytes) establish the beginning of an instruction, the disassembler will start producing the wrong code. The difficulty is that x86 instructions are varying in length, and in a generally short period the reversing tool will be synchronized again with the correct instructions.

Original Disassembly		
c745d401000000	mov	DWORD PTR [rbp-0x2c], 0x1
48c7c040a93c00	mov	rax, 0x3ca940
488d8024740300	lea	rax, [rax+0x37424]
ffe0	jmp	rax
Addition of 0x05		
05c745d401	add	eax, 0x1d445c7
0000	add	BYTE PTR [rax], al
0048c7	add	BYTE PTR [rax-0x39], cl
c073a93c	rol	BYTE PTR [rax-0x57], 0x3c
00488d	add	BYTE PTR [rax-0x73], cl
80247403	and	BYTE PTR [rsp+rsi*2], 0x3
00ff	add	bh, bh
e01d	.byte	0xe0

Figure 1: Confusing a reverse assembler by insertion of one byte of "junk"; the trailing 0xe0 will become a "loop" instruction with the next byte.

A related technique used to fool tree-based reverse engineering tools is to include a jump/branch instruction which has the address of the junk byte as the target of the branch. This will add this destination address to the list of code areas that the approach needs to work on, thus eventually forcing the disassembly to start at the wrong location. In Figure 1 we would include code that jumps to the address of the added 0x05, but with the caveat that the jump will never actually be taken when the program runs. This can easily be accomplished with an opaque predate – a small bit of code generating the results of a comparison, with the understanding that the comparison always results in a false (or true) result. This will be followed by a conditional jump based on the condition being true (or false).

A variation of junk insertion is code overlapping, as described in JLH and elsewhere. Code overlapping creates instructions that have their bytes in two (or more) other instructions. It is possible to create or select instructions with certain properties such that starting a block from two different addresses will execute two completely different sequences of code. By making sure that all instructions in the two (or more) paths always overlap, the instructions will not resynchronize until the end of the overlapped block.

3.2 Background: Anti-Disassembly

Many cybersecurity practitioners will be familiar with the x86 instruction with the operation code 0x90; it is a "no-operation" or "NOP". What may be less familiar is that this instruction exists in many forms; the one-byte form is just an easy one to remember. However, the byte sequence 0x66, 0x90 is also valid. The 0x66 is an "operand-size override prefix". This prefix indicates that the normal operand size for the instruction is not correct in this case. So, the two-byte sequence says, "override the normal operand size, for an instruction that does nothing". Other prefix bytes can also be prepended on x86 instructions, such as segment overrides; in fact, one can construct a 15-byte (the maximum allowed) sequence of bytes which form a valid instruction, which when executed does nothing. This knowledge is key to the approach by JLH.

Jämthagen uses a 9-byte NOP instruction. By formatting the operation code correctly, this no-operation instruction will contain five bytes, made up of an addressing mode and a memory address. Since the instruction itself is a no-operation the actual values in these five bytes are irrelevant since the memory address will never be accessed. This is similar to an approach where the contents of 64-bit immediate operands are used to hide executable instructions (Mahoney 2018). That method proves to be interesting in theory but not practical, due to the low likelihood of encountering 8-byte operands in actual typical programs (Mullins 2020), (Mullins 2022).

3.3 Overlapping

In order to take advantage of the 9-byte NOP and overlap instructions, Jämthagen posits the following requirements:

No alignment: the instructions must overlap each other and must never be aligned such that two instructions end at the same byte.

Both valid: each execution path – both the normal and hidden paths – must consist of valid instructions.

Both executable: each execution path must consist of executable instructions.

Recall the previous acronyms: we are overlapping a hidden execution path (HEP) with a main execution path (MEP). Requirement one is to avoid the case where a reverse assembler will re-synchronize with the hidden instruction stream part way through a block; this implies that all instructions in HEP are disjoint from those in MEP. Requirement two assures that the reverse engineering tool will not stop on an illegal instruction in the middle of the block, thus accidentally alerting an analyst that "something's wrong with this picture". This requirement is mainly concerned with MEP, as presumably HEP was hand-written in advance. Requirement three is the most difficult – if the code is executed, both MEP and HEP must execute and not generate – for example – an illegal memory reference, faulting the program. Requirement three is what forces the JLH method to use NOP instructions, as this requirement is primarily concerning MEP.

The work by Jämthagen considers instructions in MEP that will take care of obfuscating MEP: "… the exact effects the MEP has are not important since its primary function is to hide the HEP". Their idea is to then select "instrucitons which have as many "bytes that can be arbitrarily chosen as possible".

These instructions are divided into three sections, which are then referred to as "XXYYZZ". The idea is that the portion represented by "XX" contains the instruction prefixes as well as the actual operation code for the instruction. The middle "YY" section contains bytes whose actual value is not critical. These bytes might make up a memory address (which will not be accessed) or an immediate operand of an instruction. Lastly, "ZZ" needs to also be capable of having any byte assigned to it, with the note that the combination of "ZZ" and "XX" of the next instruction should decode to a valid instruction, thus satisfying requirement one. This combination of "ZZXX" is referred to as a "wrapping instruction". The HEP is embedded in a series of 9-byte NOP instructions, with the "ZZ" portion of each MEP instruction overlapping the NOP. The authors present the following example; here the top of Figure 2 shows the 9-byte NOP instructions which will hide the HEP instructions. But if instruction execution starts four bytes into the first block, at the 0x66, the hidden path is executed, as shown in the bottom of the figure.

Original Disassembly – MEP		
660f1f8466bb0100a9	nop	word ptr [rsi – 0x56fffe45]
660f1f8431c040cd80	nop	word ptr [rcx + rsi – 0x7f32bf40]
Executing From Byte Offset Four – HEP		
66bb0100	mov	bx,0x1
a9660f1f84	test	eax,0x841f0f66
31c0	xor	eax,eax
40	inc	eax
cd80	int	0x80

Figure 2: Main Execution Path (top), Hidden Execution Path (bottom)

Much of the remainder of the Jämthagen approach is concerned with improving the "look" of the MEP. For example, if a reverse engineering expert sees a long string of 9-byte NOP instructions they will start to be suspicious. An approach (which we will see again in section IV-a) is to scatter the instructions in HEP a bit farther apart and include at the end of each HEP instruction a jump to the next instruction. Although we will use the same idea, note that there are two drawbacks to this approach. First, including a jump takes away two bytes from the "payload" of the NOP instruction and pressures the author of the HEP instructions to select instructions which are very small in terms of byte count. Second, the 2-byte format of an x86 jump contains a byte displacement, limiting the distance to jumping from -128 to +127 bytes away. The JLH paper spends some time covering how to write short sections of code with these limited-length instructions. Our approach is to hand the compiler the "-Os" (optimize for size) command line option and see how things come out!

4. Materials and Methods

In this section we introduce a new methodology to see if we can improve upon the JLH method, while at the same time still holding to the three restrictions outlined in their paper. As mentioned above, previous work included an approach whereby HEP is encoded in the 8-byte operands of 64-bit instructions. The procedure works, but like in JLH, the output for MEP "looks suspicious" since there is a need for many 64-bit immediate

operands or addresses. Since our goal is to interleave instructions, we call the process "I^2". The I^2 method consists of two main phases, a placement phase and a fill phase.

4.1 Instruction Placement

The first step in the process is to craft the code which will be in the hidden execution path (HEP). Here it is preferable to have instructions that are small. The algorithm will insert jump instructions and the displacement of the jump is limited; short instructions make for close jumps. There are also a few assumptions made with respect to the code placed in HEP. Jämthagen does not specify or indicate the surrounding context for the block of code. Here we assume that our method will produce a stand-alone function for MEP which includes HEP; the function will have a standard prologue and will jump over MEP to the epilogue, thus loosening requirement number three, above. This jump is currently unconditional but can easily be modified to use an opaque predicate to ensure that instructions in MEP that may cause a fault will not be executed. In fact this predicate might rely on a parameter passed to the function, making it more difficult for one to realize that it is always the same conditional result.

The instructions in HEP are initially (manually) placed into a data structure which is built from an assembly "listing" file, typically generated by the Gnu assembler "as". For example, consider the example HEP code outlined in Jämthagen:

```
MOV         BX, 0x0001
TEST        EAX, 0x841F0F66
XOR         EAX, EAX
INC         EAX
INT         0x80
```

Figure 3: Original Test Case, from Jämthagen et al.

The data structure corresponding to this is in Figure 4. Note that there is no need for us to include the second instruction since it serves no purpose.

```
Struct mal_inst decoder[] = {
{ 0,4,no_jump,1,1, { 0x66, 0xbb, 0x01, 0x00 }, "MOV BX,0x0001" },
{ 0,2,no_jump,2,2, { 0x31, 0xc0 }, "XOR EAX,EAX" },
{ 0,1,no_jump,3,3, { 0x40 }, "INC EAX" },
{ 0,2,no_jump,0,0, { 0xcd, 0x80 }, "INT 0x80" }
};
```

Figure 4: Data Structure for I^2 using Jämthagen et al.

The fields in this structure include the location where the instruction will be placed, the number of bytes, the type of instruction as well as the number of the next instruction, the actual opcode bytes, and the original string which is used for debugging. Thus, each instruction includes meta-data with the actual bytes necessary for the instruction, as well as other information including the instruction type. This type is either a "regular" instruction, a conditional jump, or an absolute jump. (Their example does not include any jumps.) The absolute jumps are limited to 2-byte jumps with the second byte as the offset, as per x86. In the case of conditional jumps, the structure maintains the number of the next instruction assuming the condition is met, as well as the next instruction number if the condition is not met.

Using this data structure, the instructions are scattered in a random order into a "field" of bytes. The field has been initialized with random valued bytes, so the location of each instruction in HEP is tracked. Each regular opcode or conditional opcode requires an additional two bytes; these two bytes will encode an unconditional jump with the offset, and this jump will go to the next instruction in the original sequence. The instructions will not be in the field in the same order that they appear in the assembly code, and the displacements in the jump instructions are set so that the proper sequence is maintained.

The field of bytes might appear as follows, using a mixture of notations to better convey the idea. In this Figure 5, "Random" is a span of two yet unused bytes prior to the first instruction in the field, and between all subsequent instructions. The ordering of the instructions is dictated by the seed for a pseudo-random number generator, which can be set by the user. It is of course necessary to make a note of the correct entry point (the

"MOV" instruction) as this is the instruction to jump to in order to execute the HEP. If one were to write this in assembly language from scratch, it might look like Figure 5 below.

Since the data is an array of bytes there are no actual assembly language labels available; as each instruction includes a jump to the next instruction in sequence, the next step is to insert the correct 8-bit displacement into this jump instruction such that the normal instruction flow remains intact. Unfortunately, there is the potential that the instruction sequence is ordered such that the next instruction is "too far away" for this displacement. In this case since the algorithm discards this attempt and selects another random ordering of the instructions. Assuming this placement can be ordered, I^2 proceeds to the next stage.

Random bytes		
L1:	XOR	EAX, EAX
	JP	L2
Random bytes		
Entry:	MOV	BX, 0x0001
	JP	L1
Random bytes		
L3:	INT	0x80
Random bytes		
L2:	INC	EAX
	JP	L3
Random bytes		

Figure 5: Original Test Case Expanded Into Byte Field

4.2 Filling MEP

The next stage consists of randomly generating instructions to cover up those in HEP, while still satisfying the "No alignment" and "Both valid" restrictions. This is accomplished with a recursive depth-first algorithm, much like the search strategy of the language Prolog. The search function is called with a state, including the partially filled field as well as the current position. Several different events may now happen:

1. The current position may match with the beginning of an instruction in HEP.
2. The current position may be in the middle of an instruction in HEP.
3. The current position may be in the gap between HEP instructions.

Let us consider each case in turn, starting with the first. If the position corresponds to the location in the field of any instruction in HEP, this invalidates Jämthagen's "No alignment" condition; this is a failed state, and it is necessary to backtrack.

In case two, if we can decode a valid instruction from this point, we should advance to a new state. This new starting point will be the current location plus the length of the decoded instruction.

Case three is more complex. The current position is between instructions in HEP so the opportunity exists to have more flexibility. Recall from above that a technique for executable steganography is the addition of a byte of junk. Here values from 0..255 (0x00..0xff) are selected from a random permutation and placed at this location. The location is tested to see if a valid instruction results, and if so, we advance to a new state starting at the end of this new instruction. At any point in this search, the process may fail and return to a prior state. Here we may try all 256 junk bytes, with none of them yielding a valid operation code. When this happens, the process backtracks to a prior state, which may at that point try other values.

To use terminology from JLH, in this third case we are randomly selecting "XX" in the gaps between HEP and allowing "YY" to be the operation codes in the HEP instruction.

4.3 Finalizing the Sequences

Next, recall the three conditions outlined by JLH. "No alignment" states that the instructions must overlap. Since the starting point for each instruction in MEP is either inside a HEP instruction or in the gap between them, this condition is met.

"Both valid" states that both HEP and MEP are made up of legal (decodable) instructions. Those in HEP are valid from the onset. At any point in the I^2 algorithm if an instruction cannot be decoded at a certain position the search backtracks. It is true that the search may never find a valid sequence for MEP; but if a sequence is found it will be valid instructions, so this restriction is met by the algorithm.

The third condition, "Both executable", is the most difficult. As is pointed out (extensively) in JLH, certain instructions are safe and certain instructions are not safe. "Load effective address", for example, puts an address into a register but does not actually access the memory at that address. This is safe. Certain instructions are – by their very nature – not safe. Consider "LIDT" which loads the address of the interrupt dispatch table. This instruction can only be run in privilege level zero and will otherwise fault the program. Instructions which access memory are off limits since the address used will be arbitrary. Instructions that require any escalated privilege level are also bad since they will fault. But recall that here we are simply jumping over the MEP instructions, so illegal instructions are not executed. This condition is met by simply inserting a false predicate to avoid the code.

5. Discussion and Example Results

Next, we take several examples to demonstrate the operation of the method. The goal is to produce code which will, when disassembled, appear as just another part of the program. (One can argue as to whether a sequence of very strange instructions "stands out", but one can also argue that a long sequence of 9-byte NOP instructions stands out just as much.) The aim is to prevent HEP from appearing, unless the reverse engineer knows where to look and what to do.

We will examine the binary code artifacts with two tools, the popular IdaPro (HexRays 2022) reverse engineering tool and the open source Ghidra (2022) tool.

5.1 Repeating Jämthagen

We start first with the same example described in the Jämthagen paper, as shown in Figures 3, 4, and 5 above. This code is designed to run on a 32-bit Linux system, as interrupt 0x80 is a system call. In this case the system call will cause a program exit.

When this code is used for our algorithm, IdaPro shows only the function prolog and epilog, as in Figure 6:

```
; =============== S U B R O U T I N E =======================================

; Attributes: bp-based frame

sub_1199        proc near               ; CODE XREF: .text:0000000000001158↑p
                push    rbp
                mov     rbp, rsp
                jmp     short loc_11BA
; ---------------------------------------------------------------------------
                db 0C8h
                dq 403DEC9D0FEB80CDh, 0F7EBC0318183F6EBh, 0EB0001BB663B25D4h
                db 0F3h, 0F4h
; ---------------------------------------------------------------------------

loc_11BA:                               ; CODE XREF: sub_1199+4↑j
                pop     rbp
                retn
sub_1199        endp
```

Figure 6: Original Test Case, from Jämthagen et al.

Assuming the reverse engineer notices the constants within the function, the next step would be to manually instruct IdaPro to start disassembly at the first non-decoded byte. This results in uncovering more of MEP as in the following Figure 7 below.

Note that although this appears to be a strange sequence of instructions, it decodes with no illegal operations and still hides the actual aim of the instructions above. While not shown here, the Ghidra tool gives very similar results.

```
; =============== S U B R O U T I N E =======================================

; Attributes: bp-based frame

sub_1199        proc near               ; CODE XREF: .text:0000000000001158↑p
                push    rbp
                mov     rbp, rsp
                jmp     short loc_11BA
; ---------------------------------------------------------------------------
                enter   0FFFFFFFFFFFF80CDh, 0EBh
                setnl   ah
                cmp     eax, 83F6EB40h
                xor     dword ptr [rcx], 0D4F7EBC0h
                and     eax, 1BB663Bh
                add     bl, ch
                rep hlt
; ---------------------------------------------------------------------------

loc_11BA:                               ; CODE XREF: sub_1199+4↑j
                pop     rbp
                retn
sub_1199        endp
```

Figure 7: Original Test Case, from Jämthagen et al., additional instruction decoding

5.2 A More Serious Example

In terms of malware or nefarious deeds, code that simply exits a program, as above, is not too serious. We next test the following sequence of x86-64 instructions as shown in figure 8.

```
XOR             EAX, EAX
MOV             AL, 0xA9
MOV             EDI, 0xFEE1DEAD
MOV             ESI, 0x28121969
MOV             EDX, 0x4321FEDC
XOR             R10, R10
SYSCALL
```

Figure 8: Malware Testcase, Halt System

This is a Linux 64-bit operating system call and, assuming it is run as root, will immediately halt and power off the computer. (This, as a side note, makes testing and debugging occasionally problematic.) Obfuscating this code using our algorithm results in the following view in Ghidra. This tool has missed the hidden instructions entirely, as in the following (partial) Figure 9:

```
                        *************************************************************
                        *                          FUNCTION                         *
                        *************************************************************
                        undefined FUN_00101199()
          undefined         AL:1           <RETURN>
                        FUN_00101199                                XREF[1]:     00101158(c)
     00101199 55            PUSH        RBP
     0010119a 48 89 e5      MOV         RBP,RSP
     0010119d eb 45         JMP         LAB_001011e4
     0010119f 03            ??          03h
     001011a0 72            ??          72h    r
     001011a1 a5            ??          A5h
     001011a2 e9            ??          E9h
     001011a3 c3            ??          C3h
     001011a4 eb            ??          EBh
     001011a5 04            ??          04h
```

Figure 9: Malware Testcase, Halt System, Ghidra View

Manually instructing Ghidra to convert this to code results in uncovering only the main execution path as shown in figure 10 below:

```
                        undefined FUN_00101199()
          undefined         AL:1           <RETURN>
                        FUN_00101199                                    XREF[1]:     00101158(c)
     00101199 55              PUSH        RBP
     0010119a 48 89 e5        MOV         RBP,RSP
     0010119d eb 45           JMP         LAB_001011e4
     0010119f 03 72 a5        ADD         ESI,dword ptr [RDX + -0x5b]
     001011a2 e9 c3 eb        JMP         LAB_ffffffffe614fd6a
              04 e6
     001011a7 57              PUSH        RDI
     001011a8 a0 b8 31        MOV         AL,[DAT_bab1cf15ebc031b8]
              c0 eb 15
              cf b1 ba
     001011b1 dc fe           FDIV        ST6,ST0
     001011b3 21 43 eb        AND         dword ptr [RBX + -0x15],EAX
     001011b6 05 c7 ea        ADD         EAX,0xff4eac7
              f4 0f
     001011bb 69 4d 31        IMUL        ECX,dword ptr [RBP + 0x31],0x770aebd2
              d2 eb 0a 77
     001011c2 34 b0           XOR         AL,0xb0
     001011c4 a9 eb 0a        TEST        EAX,0xa1870aeb
              87 a1
     001011c9 98              CWDE
     001011ca 69 0f 05        IMUL        ECX,dword ptr [RDI],0x39d4eb05
              eb d4 39
     001011d0 81 bf ad        CMP         dword ptr [RDI + -0x11e2153],0xadb005eb
              de e1 fe
              eb 05 b0 ad
```

Figure 10: Malware Testcase, Halt System, Ghidra View, Manual Conversion

Lastly, we consider a case where a nefarious malware author may wish to hide an encryption algorithm of some sort. Consider the following 'C' function to do a simple block copy with an exclusive-or:

```
void decrypt( unsigned char *a, unsigned char *b, unsigned n )
{
    for( unsigned I = 0; i < n; i++ )
        *b++ = *a++ ^ 0x42;
}
```

Figure 11: Simple 'C' Block Decryption Function – XOR with 0x42

This code is compiled using GCC and optimized for space, generating an assembly function with nine instructions, which is then inserted as the hidden execution path. The following figure shows the (partial) result in IdaPro:

```
sub_11AF        proc near               ; DATA XREF: .text:0000000000001145↑o

; FUNCTION CHUNK AT .text:000000000000121F SIZE 00000046 BYTES

                push    rbp
                mov     rbp, rsp
                jmp     short loc_11FB
; ---------------------------------------------------------------------------
                insd
                insb
                movsb
                imul    ecx, [rdx+22EB070Ch], 24h
                imul    esi, [rsi+31h], 0B1CFF5EBh
                xor     [rsi-14F9F378h], bh
                add     al, bh
                adc     eax, 0EBC0FF48h
                adc     eax, 77E58F20h
                cmp     dword ptr [rcx], 87E3EBC2h
                xchg    eax, edi
                or      eax, 0EB42F183h
```

Figure 12: Malware Testcase, Decoder Function in IdaPro

It appears here that IdaPro has again defaulted to the simple view of performing a linear disassembly over all of the main execution path, never seeing any of the hidden instructions. Calling externally to this function address plus (in this case) 0x3E causes the decoder to execute. Ghidra and other tools show the same view of this code.

6. Conclusions and Future Work

The methods outlined by Jämthagen, Lantz, and Hell do an excellent job of hiding executable instructions inside of a series of no-operation instructions. The advantage of their method lies in the fact that one can still safety execute the main instruction sequence with no ill effects. However, this severely limits the available

opportunities to hide instructions; providing an appropriate opaque predicate to avoid executing the main path opens better opportunities for protecting the secret steganographic instruction payload.

The method discussed here opens up greater opportunities in terms of instruction hiding. First, it removes the restriction that the main execution path consist of no-operation instructions. Our argument is that a long string of these NOP instructions may stand out when reverse engineering, and if the sequence is already "somewhat odd" it doesn't matter which instructions make it look odd. Second, if one opens the possibility of using any hiding instruction, as opposed to only NOP instructions, those instructions which are hidden within are no longer restricted. As a result, longer code sequences (relative to the original work) can be hidden and still execute correctly.

Looking forward, the method outlined here utilizes the "xed" (2022) package to determine whether the bytes at a certain location in our "field" decode to an instruction. Potential future research is to restrict the decoded instructions to ones that use only immediate operands, for example, to make the main execution path appear more like normal code. Alternatively, one can construct, using "xed", instructions with certain characteristics, and have the package return back the actual opcode bytes. These opcode bytes could be used instead of the random selection method currently used by the algorithm.

References

AMD, *AMD64 Architecture Programmer's Manual— Volume 3: General-Purpose and System Instructions*, [online], https://www.amd.com/system/files/TechDocs/24594.pdf, accessed November 2022.

Binary Ninja, [online], https://binary.ninja, accessed November 2022

Defuse, [online], https://defuse.ca/online-x86-assembler.htm, accessed November 2022.

Ghidra, [online], https://ghidra-sre.org, accessed November 2022.

HexRays IdaPro, [online], http://www.hex-rays.com/products/ida/index.shtml, accessed November 2022.

Hamid, N., Yahya, A., Ahmad, R. B., & Al-Qershi, O. M. (2012). Image steganography techniques: an overview. *International Journal of Computer Science and Security (IJCSS)*, *6*(3), 168-187.

Hopper, [online], https://www.hopperapp.com, accessed November 2022.

Intel (2006). *Intel® 64 and IA-32 Architectures Software Developer's Manual Volume 2A: Instruction Set Reference, A-M*, 2006.

Jamthagen, C, Lantz, P. and Hell, M. (2013). "*A new instruction overlapping technique for anti-disassembly and obfuscation of x86 binaries*", 2013 Workshop on Anti-malware Testing Research (WATeR), 978-1-4799-2476-9, IEEE.

Linn, C. and Debray, C. (2003). "*Obfuscation of Executable Code to Improve Resistance to Static Disassembly*", Proceedings of the ACM Conference on Computer and Communications Security (CCS 2003), Washington, DC, USA, Oct. 27-30, 2003.

LinuxDevCenter, [online], http://www.linuxdevcenter.com/cmd/cmd.csp?path=o/objdump, accessed November 2022.

Mahoney, W. (2015) "*Modifications to GCC for Increased Software Privacy*", International Journal of Information and Computer Security, Vol. 7, No. 2, 3, 4.

Mahoney, W., Franco, J., Hoff, G. and McDonald, J. T. (2018) "*Leave it to Weaver*", 8th Software Security, Protection, and Reverse Engineering Workshop, Puerto Rico, December 3-4.

Mahoney, W., McDonald, J. T., [online], "*Enumerating x86-64 – It's Not as Easy as Counting*", https://www.unomaha.edu/college-of-information-science-and-technology/research-labs/_files/enumerating-x86-64-instructions.pdf, accessed November 2022.

Morkel, T., Eloff, J. H., & Olivier, M. S. (2005). "An overview of image steganography", In Information Security South Africa (ISSA) (Vol. 1, No. 2, pp. 1-11)

Mullins, J. A., McDonald, J. T., Mahoney, W. and Andel, T. (2020) "*Evaluating Security of Executable Steganography for Digital Software Watermarking*", 2020 Cybersecurity Symposium, Moscow, Idaho.

Mullins, J. A., McDonald, J. T., Mahoney, W. and Andel, T. (2022) "*Evaluating Security of Executable Steganography for Digital Software Watermarking*", IEEE SoutheastCon 2022, Mobile, AL, March 31 – April 3, 2022.

OllyDbg, [online], http://www.ollydbg.de/, accessed November 2022.

Pearce, W., (2008). *Reverse Engineering Code with IDA Pro*. Elsevier Science.

Relyze, [online], https://www.relyze.com/overview.html, accessed November 2022.

Shinn, S. and Mahoney, W. (2011) "*Optimal Values for Disrupting x86-64 Reverse Assemblers*", International Journal of Computer Science and Network Security, Volume 11, Number 11.

Udis86, [online], http://udis86.sourceforge.net/, accessed November 2022.

x64dbg, [online], https://x64dbg.com, accessed November 2022.

XED, [online], https://intelxed.github.io/ref-manual/, accessed November 2022.

Developing Privacy Incident Responses to Combat Information Warfare

Sean McElroy and Lisa McKee
Dakota State University, Madison, USA

me@seanmcelroy.com
lisa.mckee@trojans.dsu.edu

Abstract: Violations of privacy harm real people, and as nation-state actors grow their information warfare capabilities, civilians suffer these harms as part of coordinated and targeted actions on objectives. When privacy harms manifest, they allow threat actors to injure data subjects by weaponizing their information to harm individuals, communities, and societies. These attacks injure civilians as the confidence of legitimate authorities, institutions, and defences is eroded, and consequences may impact national security. Distinct from cybersecurity, privacy depends upon confidentiality, integrity, and availability but encompasses a unique set of concerns. Whereas security incident response has an established practice and research history, approaches to privacy incident response, such as unauthorized disclosure, are not well researched or documented in academic literature in the unique context of privacy. By mapping privacy harm to techniques and tactics, a cohesive framework emerges to distinguish tailored mitigation strategies for each. This paper proposes a conceptual model and classification framework for privacy-related harms, tactics, techniques, and mitigation strategies to address sophisticated privacy threat actors. Using this model and framework, contingency planners can develop privacy incident response strategies to defend against the privacy harms of information warfare.

Keywords: incident response, data privacy, information warfare, disinformation

1. Introduction

Attacks on data privacy can have a paralyzing effect on a civilian population. However, while information security response strategies are well established in the literature, privacy is often only considered a secondary concern in a security incident. While many detective and preventative security controls are rooted in technical tooling and implementation, security practitioners often lack privacy awareness, privacy engineering skills, or appropriate conceptual models and action frameworks that help them distinguish the unique components of privacy incident preparation and response.

Security incident response plans partly address many privacy harms on a frequency basis, as the highest number of privacy incidents involve a data breach caused by a loss of security. However, many forms of privacy harm, both individual and societal, are not caused by a security incident and may have no technological basis for remediation. Security incident response is an insufficient mechanism for addressing privacy harm. An approach to address privacy by adding it to existing security approaches is incomplete in addressing privacy harms that do not result from data breaches. Distorted propaganda, disinformation, and incitement, for instance, may manifest as a national security concern or, when targeted at a company or individual, substantial economic or personal harm. This research proposes a conceptual model for privacy incident response that addresses the unique domain of privacy and a response framework to address the different harms enumerated in Solove's *A Taxonomy of Privacy*.

2. Literature Review

Privacy harms are distinct from information security consequences. However, the development of models and frameworks expressive enough to capture both is stymied by a false equivalency drawn between security and privacy in fundamental industry standards and practices. For example, the NIST Privacy Framework specifies response controls, but only in the context of shared "cybersecurity-related privacy events" (National Institute of Standards and Technology, 2020, p. 7), while simultaneously acknowledging that failing to address societal risks leads to an incomplete privacy risk management approach. Privacy harm to even individuals can have significant national security consequences, as the rise of widely deployed biometric profiling technology and leaks of highly sensitive personal data included in the 2015 Office of Personnel Management data breach have both harmed US intelligence missions and the country's ability to maintain and gather foreign intelligence partners. (Dorfman, 2020) On a larger scale, societal privacy harms can result in real-world dangers, as their manifestation can "demotivate [military] personnel or create unrest in the [civil] population" (Fritsch and Fischer-Hübner, 2018).

Existing academic literature attempts to identify appropriate models, frameworks, and processes to address privacy incident response. Abdullah proposes a modified conceptual framework incorporating the Forrester GRC

playbook (Abdullah, 2018), but data breaches are the only harm identified for a response. Abdullah references NIST special publication 800–61, NIST special publication 800–122, and the Nymity Privacy Management Accountability Framework as sources for further development of an incident response framework. Both 800-61 and 800-121 detail incident responses from the perspective of a technologist that is tackling a loss of confidentiality (McCallister, Grance and Scarfone, 2010; Cichonski et al., 2012), a principle of information security (McCumber, 2005) that is not inclusive or comprehensive of privacy principles. The Nymity framework includes critical operational processes that may prepare for or prevent non-breach-related privacy incidents; however, the actual incident response activities only cover data breaches (Nymity Inc., 2017). The tendency for industry best practices and guidance as well as academic models tend to address data breaches or unauthorized access resulting in privacy harms in the form of disclosure, but they do not provide a model or framework to address other forms of privacy harm, particularly those instigated by nation-state actors in long-term adversarial campaigns.

Solove's A Taxonomy of Privacy is an informative reference that outlines sixteen unique privacy harms to examine what privacy incidents other conceptual models and incident response frameworks fail to address. (Solove, 2006) Of these sixteen, the literature referenced above addresses at most two of these from Solove's taxonomy: breach of confidentiality and disclosure. Indeed, these are arguably the two most costly and prevalent forms of privacy harm for the industry, with the per-record cost of data breaches hitting an all-time high of $164 in 2022. (Ponemon Institute, 2022)

Data subjects do not experience privacy harm universally in the same way or with the same relative effect. Varying values, beliefs, and attitudes on privacy are not easily classified cross-culturally on geographic or political boundaries (Zabihzadeh et al., 2019). While Solove acknowledges this in his seminal taxonomy (Solove, 2006), the differences in how cultures value privacy and experience subsequent harms necessitate different privacy incident responses commensurate with mitigating the harm that has actually occurred to the extent it affects data subjects. Certainly, harm can be statutorily experienced by data processors and controllers in the form of punitive damages or fines, as prescribed by the EU's General Data Protection Regulation. However, the relative harm experienced by a data subject may vary by cultural, locale, and personal factors. For example, the psychological harm a privacy violation causes the data subject, which research shows is more significant among consumers residing in countries without privacy regulations (Bellman et al., 2004). This result is more significant when harms are perceived to be reputational and cause an economic future impact, such as difficulties in finding housing or employment (Anabo, Elexpuru-Albizuri and Villardón-Gallego, 2019). Blackmail, harm enumerated by Solove, may require different mitigations in a privacy incident response plan for data subjects that experience societal impacts disproportionately to data subjects in social networks that do not value reputation or suggestions of impropriety the same (Al-Saggaf, 2016).

By contrast, Decisional Interference is difficult to both identify and quantify. For example, the Mueller report on Russian interference in the 2016 United States presidential election concluded it occurred "in a sweeping and systematic fashion" (Mueller III, 2019). Subsequently, trust in elective democratic processes in America is significantly declining (Rainie, Keeter and Perrin, 2019). As quantitative macroeconomic research has concluded that lower levels of trust result in lower national economic performance and future growth (Mularska-Kucharek and Brzeziński, 2017), the costs of privacy harm other than data breaches may be both enormously underappreciated as well as lacking commensurate incident response models to protect national security interests. West acknowledges this supposition and further posits additional complications caused by the rise of encrypted communications channels that allow adversaries to conduct disinformation campaigns outside of the reach of public Internet monitoring and community analysis. (West, 2021) The rise of highly automated and expansive disinformation bots on social media coupled with direct messaging capabilities and tailored profiling of the populace may result in significant blackmail, extortion, distortion, and decisional interference Solovian privacy harms materializing. While one may think harms, such as Decisional Interference, as uniquely targeted at a single individual, big data, encryption, and cloud infrastructure may allow these to happen quickly at a vast scale, outside the visibility of intelligence and policing forces.

Returning to the subject of "trust" in the literature, repairing or rebuilding it is the subject of some data-breach-centric incident response frameworks. It is a unique characteristic not present in security incident response frameworks. Wan and Zhang noted the value of trust and analysed existing methods commercial entities use to attempt to increase it. Their analysis revealed that "apology" and "providing information", in the form of answering questions about a privacy breach, were the two most common methods used, with victim monetary or credit monitoring compensation or management action plans as the final two methods observed. (Wan and Zhang, 2014) The efficacy of these methods was not assessed, and it stands to reason that even if they

succeeded in reducing economic harm and increasing trust, they are mainly inapplicable to privacy harms in Solove's taxonomy unrelated to disclosure.

3. Methodology

This paper proposes a conceptual model and outlines a privacy incident response framework using a design science research methodology using the engineering lifecycle (Wieringa, 2014). Design science research allows for developing innovative artefacts to obtain a goal (Simon, 1988; Vaishnavi, 2007). It will require multiple iterations of model design and review with industry professionals and colleagues. The research will include interviews with industry experts to get their input and responses to specific questions to the research question. This question is how can privacy incident response be developed and refined with a unique ontology and approach that best addresses each privacy harm, as identified by Solove.

With a suitable conceptual model developed to approach the question, we propose a preliminary framework that maps mitigations to tactics ("high-level descriptions of behaviour") and techniques ("detailed descriptions of behaviour in the context of a tactic"), as both terms are defined by NIST (Johnson *et al.*, 2016). By leveraging a common vocabulary, the proposed framework appeals to the familiarity of threat modelling and incident response, which are well-developed concepts in information security. It establishes the unique approaches required to build privacy incident response programs.

4. Conceptual Model

Solove frames privacy harms by the underlying activity that creates the issue, culminating in a taxonomy detailing four primary classes: information collection, information processing, information dissemination, and invasion (Solove, 2006, pp. 483–490). This framework is helpful because it focuses on activities to which mitigating responses can be directly mapped. It is also a useful starting point as it focuses, in the first three, on flows into, within, and out of data holders. Privacy incident response must be implemented, at least in part, by data holders, even if at the direction of a national data protection authority.

Consider, however, that the casual elements that create harm may be intentional or accidental. Whether a specific privacy violation is legal in any given jurisdiction, that it is in any suggests intentional privacy threat actors possess *mens rea* or a culpable mental state. American law suggests four types, in descending order of culpability: "purpose, knowledge, recklessness, and negligence" (Kagan, 2021). While this research does not intend to provide a response framework that adheres to any specific legal framework, considering the diverse ways intentional privacy harm can occur is essential to model privacy threats and build a comprehensive incident response plan. Intentional actors may exhibit persistence and adaptation and be highly skilled and well-resourced if supported by a nation-state actor. The conceptual model for intentional actors causing harm is illustrated below in Figure 1. While goal formation may be less evident for reckless and negligence cases, we posit they are still discernible in a culpable actor's outcomes.

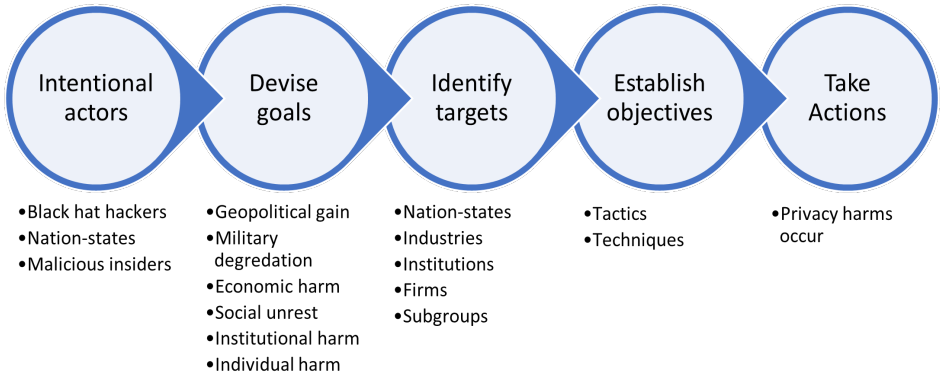


Figure 1: Conceptual Model for Intentional Privacy Harms

For accidentally created harm, we consider such accidents possible because of a failure to design a system with sufficient privacy-by-design safeguards. Cavoukian's Privacy by Design approach is further developed by Hoepman in the four technical data-oriented and process-oriented strategies (Cavoukian, 2011, 2016; Hoepman, 2020). As such, failures to minimize, separate, abstract, hide, inform, control, enforce, or demonstrate privacy in a system allow for accidental harm to occur in our conceptual model. The lack of intent does not make

accidentally created harm any less severe. While design flaws may not exhibit the same type of persistence and adaptation as intentional privacy threat actors, they may be pervasive and require considerable time and effort to design, implement, and deploy appropriate mitigations. In addition to privacy design flaws, many systems depend on security properties to deliver privacy properties. A system that correctly engineers privacy protections but has failing security controls may still result in accidental harm in this model; that is, we consider intentional attacks on security that consequently result in privacy harm within this second conceptual model. As illustrated below in Figure 2, techniques are still evident, though, without the element of intent, they represent how potential builds and not a specific planned action.

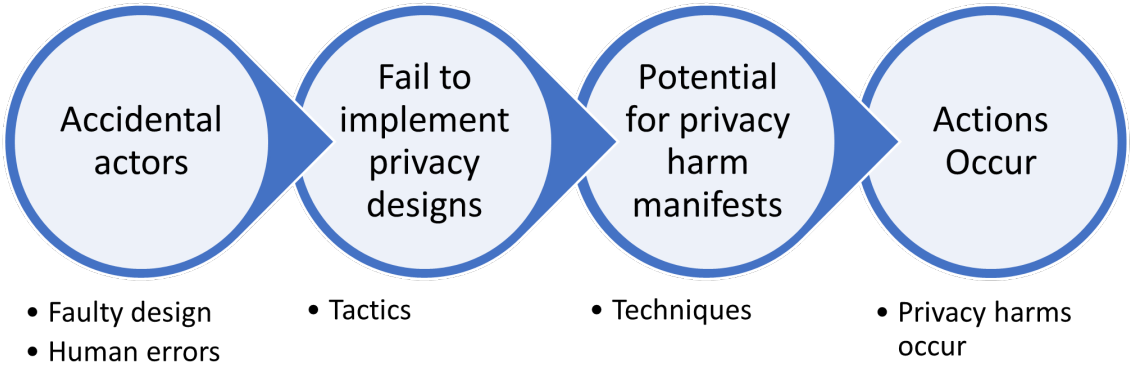


Figure 2: Conceptual Model for Accidental Privacy Harms

5. Privacy Incident Response Framework

Following the conceptual models above, a privacy incident response framework must allow for a repeatable process for incident responders to minimize privacy harm. Two popular security incident response frameworks denote similar processes in an adjacent space. The SANS Incident Response Plan enumerates six processes: preparation, identification, containment, eradication, recovery, and lessons learned (Kral, 2011; SANS Institute, 2016). By comparison, the NIST Cyber Security Framework enumerates four distinct procedural steps: (1) preparation, (2) detection and analysis, (3) containment, eradication, and recovery, and (4) post-incident activity (Cichonski *et al.*, 2012). However, unlike security incident response, where the goal is to restore the security properties of a system, the goal of privacy incident response is to minimize harm on all timescales. A Pandora's Box analogy exemplifies this fundamental difference: some privacy harms have lasting consequences that no specific activity can undo. No 'server rebuild' or gag order can restore privacy once knowledge about communication contents or associations of a data subject is disclosed to the public. Consequently, privacy incident response processes must differ to address this disparity.

5.1 Privacy Incident Response Plan (PIRP)

As discussed, privacy incident responses often do not contain harms that have already occurred, though they may contain new instances harming additional data subjects or prevent the expansion of harms (such as disclosure growing to include aggregation) from happening to the same affected subjects. For this reason, containment and eradication, terms used in security response planning, have no clear analogues in privacy incident response. Furthermore, harm recovery may take months, even years, and so privacy incident responses operate on significantly different timelines. This reality changes the nature of incident response management, resourcing, and tooling required to manage short- and long-term outcomes successfully. With this in mind, we propose the following process framework for this response, a privacy incident response plan (PIRP), illustrated below in Figure 3.

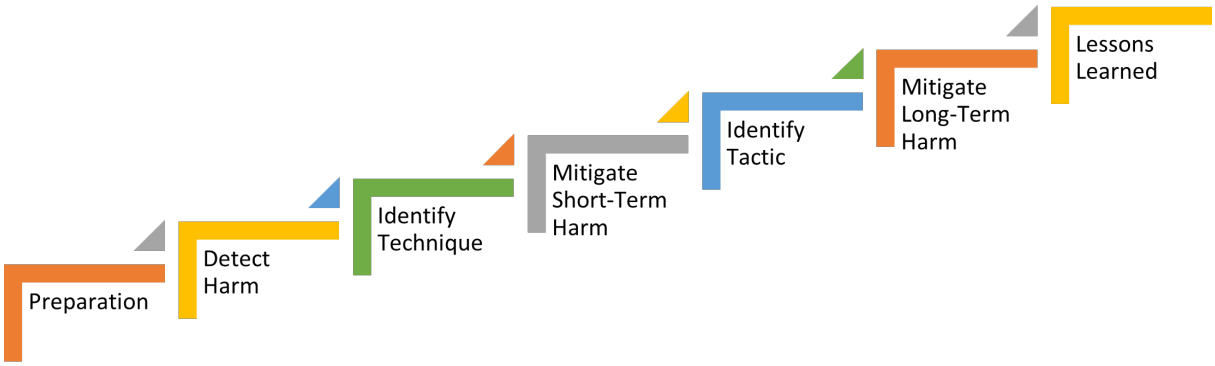


Figure 3: Privacy Incident Response Plan (PIRP) framework

5.1.1 Preparation

Preparing for using the PIRP requires knowledgeable and skilled resources who understand privacy concepts, the current legal environment, and the human impacts of specific privacy harms. PIRP plan updates and ongoing readiness training through tabletop exercises or simulated incident response scenarios prepare both incident responders and dependent legal and executive authorities to perform their duties in the face of an actual event.

5.1.2 Detect Harm

Detection mechanisms are required to initiate all subsequent plan actions. For privacy incidents, such as breach of confidentiality or unauthorized disclosure that stem from security incidents, the declaration of a security incident should also trigger a presumptive invocation of the PIRP. Other privacy-specific inputs may detect historical or ongoing harm, including direct reports from data subjects and detect leakage or use of mock or real data subject records in unexpected circumstances. Incident responders must complete all other plan steps when harm is detected and confirmed.

5.1.3 Identify Technique

Because a PIRP focuses on minimizing harm, privacy incident responders need to quickly understand the "how" before the "why" of harm occurring. An entire ontology of techniques that can cause any of the 16 Solovian privacy harms would be extensive to enumerate and describe and is outside this research's scope. However, a contingency planner could document such a technique in the scope of a nation-state PIRP, as illustrated in Table 1.

Table 1: Example technique to harm privacy

ID	Name	Description
TE0001	Disclosure Through Adjacent Queries	Adversaries may learn large amounts of private information by abusing targeted search functions in systems. *Example*: An application that allows one to find individuals on a dating website within 5 miles of a given street address could automate queries to exfiltrate the population of data subjects and pinpoint exact street addresses through triangulation of responses.

Identifying illustrative example TE0001 requires consideration in the Preparation phase of the PIRP, provided a privacy threat model determined it was a relevant privacy threat to an organization or system. Such harm detection mechanisms could be an observation of a rise in queries from a singular IP address or an anomaly in query volume that is uncharacteristic.

5.1.4 Mitigate Short-Term Harm

Individual techniques deserve specific mitigating treatments tailored to the technique. The purpose of this step is similar to 'containment' in a security incident response plan. It is to quickly stop new data subjects from being affected by this technique and continued manifestations of the technique for already-affected data subjects.

Each technique catalogued by a contingency planner and accounted for in a PIRP should have a range of mitigations. For illustrated technique TE0001, these treatments could include, as an example:

1. M0001 Rate-limit query features that return data subject object identifiers
2. M0002 Disable the query feature that returns data subject object identifiers
3. M0003 Disable access to the system providing the query feature that returns data subject object identifiers

5.1.5 Identify Tactic

Understanding the high-level descriptions of behaviour that caused privacy harm is an activity similar to root cause analysis. By viewing the specific technique that caused injury through the lens of overall behaviours, an incident responder can identify approaches for mitigating long-term harm. This is where two different conceptual models for PIRP are essential, as accidental causes are more likely related to systemic design or process problems. In contrast, intentional actors are more likely to manifest varied future harms or magnify damage through specific, additional actions.

While a comprehensive list of intentional and accidental tactics is beyond the scope of this research, an illustrative example of privacy harm tactics perpetuated by intentional actors is detailed below in Table 2.

Table 2: Example of intentional tactic to harm privacy

ID	Name	Description
TI0001	Disruptive Dissemination	The adversary is trying to harm the availability of a whole system, its components or features, or cause delays that degrade its service. *Example*: An actor scrapes data from a system and publishes an enriched version correlated with other data to cause embarrassment that provokes a reaction that causes a valuable service to be restricted or disabled. This tactic relates to the Solove harms "Aggregation" and "Increased Accessibility" for affected data subjects.

5.1.6 Mitigate Long-Term Harm

Because privacy harms cannot be contained and eradicated in the bounded way many security incidents can be, strategies for addressing long-term harms must be developed at both the technique and tactic levels in a PIRP. Considering a PIRP ontology that parented the example technique "Disclosure Through Adjacent Queries" from Table 1 under the tactic "Disruptive Dissemination" from Table 2, long-term harm mitigations exist for both the general tactic category and the specific technique. Long-term system design considerations at the tactical level might include preparing systems to operate with configurable reduced functionality to reduce the chances that disclosures through some features do not require the whole system to be disabled to remove access to a feature. At a technique level, reducing relationships between data or the visibility of specific fields through a query function may be a longer-term data transformation mitigation (Senarath and Arachchilage, 2019) for this particular vector of privacy harm.

6. Discussion

6.1 Analysis

Considering advanced persistent threat actors can and have sought to inflict privacy harm, we believe this approach sets the appropriate context for combatting related information warfare tactics. When privacy is considered only in terms of security incident response, it is addressed to the extent that security directly provides privacy. In the context of Solove's taxonomy of privacy harms, these are limited to what that research categorizes as 'Information Dissemination' harms. Existing incident response frameworks fail to consider the potential for well-resourced intentional actors who may persistently seek to cause and magnify privacy harms. Subsequently, they do not contemplate long-term harm mitigation strategies. Short-term harm mitigation strategies often include actions of limited efficacy: notice of disclosure, promises of change, and temporary consumer credit services to address consequential identity theft. These novel artefacts address these shortcomings and are conceptually complete in addressing the shortcomings of pre-existing approaches.

6.2 Limitations

This work establishes a conceptual model and privacy incident response framework design artefacts. However, these elements are illustrated but not comprehensively developed as a complete ontology. Incident responders need to model privacy threats and relevant harms and devise harm mitigation strategies suited to how they collect, use, process, and share information.

The model and the resulting framework have not yet been comprehensively tested with the design science methodology against documented privacy incidents to measure the magnitude of their impact or the performance of specific features of either artefact in the context of privacy harm mitigation.

6.3 Future Work

This new contribution of a conceptual model and privacy incident response plan framework would benefit from formal design science model testing. While the meta-requirements, meta-design, and theories are established in this paper, formal model testing that validates the use of these artefacts results in better privacy incident response outcomes, as proposed for this research methodology (Kuechler and Vaishnavi, 2008), for both short and long term harm reduction are essential areas for future work.

Techniques, tactics, and mitigation strategies are illustrated in this work. Developing a framework library repository of these artefact dimensions would significantly improve the usability of these artefacts by practitioners. For instance, a community-driven approach to soliciting feedback to develop these areas may yield significant coverage for each of the Solove privacy harms beyond what any single research group may develop. If informed by analysis of historical privacy incidents or future applications, subsequent technical action research may significantly add to the impact through iterative refinement.

Finally, operationalizing this model and framework for practitioners through developing PIRP awareness, education, and training materials could include tutorials and guides usable for those with limited experience in incident response or data privacy. Promulgating these findings through industry standards bodies, such as NIST and ANSI, may enhance existing privacy protection guidance and offer new incident response processes and procedures to existing standards.

7. Conclusions

Privacy incident response as a holistic program that addresses long-term and short-term harms using a systematic approach is not established in the literature apart from the context of security incident response or “disclosure” privacy harm. Indeed, data breaches are increasing in frequency and magnitude. However, security incident response plans often neglect the long-term effects of such disclosure, which include blackmail, extortion, distortion, and decisional interference. Using Solove’s ‘Taxonomy of Privacy’ provides a valuable framework to identify harms unaccounted for by existing incident response plans. Further, considering data privacy harms are not only caused by accidents but could be the underlying strategic objective of intentional threat actors, mapping tactics, techniques, and harms and enumerating specific short and long-term response strategies is necessary to address the lasting consequences of privacy harm. The conceptual model and privacy incident response framework presented in this design science research offer new tools that allow incident responders to address data privacy concerns directly.

References

Abdullah, H. (2018) ‘A Structural and Navigational Method for Integrated Organizational Information Privacy Protection’, in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa: IEEE, pp. 1–9. Available at: https://doi.org/10.1109/ICABCD.2018.8465138.

Al-Saggaf, Y. (2016) ‘An Exploratory Study of Attitudes Towards Privacy in Social Media and the Threat of Blackmail: The Views of a Group of Saudi Women’, *The Electronic Journal of Information Systems in Developing Countries*, 75(1), pp. 1–16. Available at: https://doi.org/10.1002/j.1681-4835.2016.tb00549.x.

Anabo, I.F., Elexpuru-Albizuri, I. and Villardón-Gallego, L. (2019) ‘Revisiting the Belmont Report’s ethical principles in internet-mediated research: perspectives from disciplinary associations in the social sciences’, *Ethics and Information Technology*, 21(2), pp. 137–149. Available at: https://doi.org/10.1007/s10676-018-9495-z.

Bellman, S. *et al.* (2004) ‘International Differences in Information Privacy Concerns: A Global Survey of Consumers’, *The Information Society*, 20(5), pp. 313–324. Available at: https://doi.org/10.1080/01972240490507956.

Cavoukian, A. (2011) 'The 7 Foundational Principles - Implementation and Mapping of Fair Information Practices'. Information and Privacy Commissioner of Ontario. Available at: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf.

Cavoukian, A. (2016) 'The Time to Embed Privacy, by Design is Now: Into IoT, AI, and Big Data'. *SAIConference*, London, England, 4 August. Available at: https://www.youtube.com/watch?v=mzk9Ijw9O_o (Accessed: 30 August 2021).

Cichonski, P. *et al.* (2012) *Computer Security Incident Handling Guide: Recommendations of the National Institute of Standards and Technology*. NIST SP 800-61r2. National Institute of Standards and Technology, p. NIST SP 800-61r2. Available at: https://doi.org/10.6028/NIST.SP.800-61r2.

Dorfman, Z. (2020) 'China Used Stolen Data to Expose CIA Operatives in Africa and Europe', *Foreign Policy*, 21 December. Available at: https://foreignpolicy.com/2020/12/21/china-stolen-us-data-exposed-cia-operatives-spy-networks/.

Fritsch, L. and Fischer-Hübner, S. (2018) 'Implications of Privacy & Security Research for the Upcoming Battlefield of Things', *Journal of Information Warfare*, 17(4), pp. 72–87.

Hoepman, J.-H. (2020) *Privacy Design Strategies (The Little Blue Book)*.

Johnson, C.S. *et al.* (2016) *Guide to Cyber Threat Information Sharing*. NIST SP 800-150. National Institute of Standards and Technology, p. NIST SP 800-150. Available at: https://doi.org/10.6028/NIST.SP.800-150.

Kagan, E. (2021) *Charles Borden, Jr. v. United States*.

Kral, P. (2011) 'Incident Handler's Handbook'. SANS Institute. Available at: https://sansorg.egnyte.com/dl/6Btqoa63at (Accessed: 1 November 2021).

Kuechler, B. and Vaishnavi, V. (2008) 'On theory development in design science research: anatomy of a research project', *European Journal of Information Systems*, 17(5), pp. 489–504. Available at: https://doi.org/10.1057/ejis.2008.40.

McCallister, E., Grance, T. and Scarfone, K.A. (2010) *Guide to protecting the confidentiality of Personally Identifiable Information (PII)*. 0 edn. NIST SP 800-122. Gaithersburg, MD: National Institute of Standards and Technology, p. NIST SP 800-122. Available at: https://doi.org/10.6028/NIST.SP.800-122.

McCumber, J. (2005) *Assessing and managing security risk in IT systems: a structured methodology*. Boca Raton, FL: Auerbach Publications.

Mueller III, R.S. (2019) *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. U.S. Department of Justice, p. 448. Available at: https://www.justice.gov/archives/sco/file/1373816/download (Accessed: 9 September 2021).

Mularska-Kucharek, M. and Brzeziński, K. (2017) 'The Economic Dimension of Social Trust', *European Spatial Research and Policy*, 23(2), pp. 83–95. Available at: https://doi.org/10.1515/esrp-2016-0012.

National Institute of Standards and Technology (2020) *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management Version 1.0*. NIST CSWP 01162020. Gaithersburg, MD: National Institute of Standards and Technology. Available at: https://doi.org/10.6028/NIST.CSWP.01162020.

Nymity Inc. (2017) 'Nymity Privacy Management Accountability Framework'. Available at: https://vvena.nl/wp-content/uploads/2018/04/PMAF-Poster.pdf (Accessed: 11 September 2022).

Ponemon Institute (2022) *Cost of a Data Breach Report 2022*. Armonk, NY: IBM Security, p. 59. Available at: https://www.ibm.com/downloads/cas/3R8N1DZJ.

Rainie, L., Keeter, S. and Perrin, A. (2019) 'Trust and Distrust in America', *Pew Research Center - U.S. Politics & Policy*, 22 July. Available at: https://www.pewresearch.org/politics/2019/07/22/trust-and-distrust-in-america/.

SANS Institute (2016) 'SANS 504-B Incident Response Cycle: Cheat Sheet'. Available at: https://www.sans.org/media/score/504-incident-response-cycle.pdf (Accessed: 19 January 2022).

Senarath, A. and Arachchilage, N.A.G. (2019) 'A data minimization model for embedding privacy into software systems', *Computers & Security*, 87, p. 101605. Available at: https://doi.org/10.1016/j.cose.2019.101605.

Simon, H.A. (1988) 'The Science of Design: Creating the Artificial', *Design Issues*, 4(1/2), p. 67. Available at: https://doi.org/10.2307/1511391.

Solove, D.J. (2006) 'A Taxonomy of Privacy', *University of Pennsylvania Law Review*, 154(3), p. 477.

Vaishnavi, V.K. (2007) *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. 1st Edition. Auerbach Publications. Available at: https://doi.org/10.1201/9781420059335.

Wan, L. and Zhang, C. (2014) 'Responses to trust repair after privacy breach incidents', *Journal of Service Science Research*, 6(2), pp. 193–224. Available at: https://doi.org/10.1007/s12927-014-0008-2.

West, L.B. (2021) 'Beyond Fighting Words: Reconceptualizing Information Warfare and its Legal Barriers', *National Security Law Journal*, 8(2), p. 162.

Wieringa, R.J. (2014) *Design science methodology for information systems and software engineering*. New York, NY: Springer Berlin Heidelberg.

Zabihzadeh, A. *et al.* (2019) 'Cultural differences in conceptual representation of "Privacy": A comparison between Iran and the United States', *The Journal of Social Psychology*, 159(4), pp. 357–370. Available at: https://doi.org/10.1080/00224545.2018.1493676.

Use of Intrusion Detection Systems in Vehicular Controller Area Networks to Preclude Remote Attacks

Anthony J Monge, Todd Andel
University of South Alabama, Mobile, Alabama, USA

ajm902@jagmail.southalabama.edu
tandel@southalabama.edu

Abstract: Security is always at the forefront of our thoughts whether we know it or not. Mindlessly, people go about their daily lives with security a part of everything performed. Is the house door locked? Is the phone off and upside down on the table so no one can see it? Is the computer at home/work logged out and secure? However, when thinking about our vehicles, the normal person stops at locking the door. Problem is that our vehicles' electronic systems are unprotected. Vehicles today are essentially personal computers with wheels. It is arguable that vehicles are incredibly safe, but that safety is an illusion. The computers that control our cars have essentially zero security in place to protect them. It is a chilling notion to have the knowledge your brakes could be rendered useless while moving at a high rate of speed. On top of that, this could be done and leave essentially no trace it had been performed. The main crux of this insecurity is the Controller Area Network (CAN) utilized by the vehicles' electronics. This paper outlines the current vulnerabilities that vex this network system and why those issues have remained unsolved. It also outlines a plausible solution to get the security community moving in the right direction. While this solution is a mere small step toward a robust network, it does alert the operator to a potential network attack. With this knowledge, the driver of the vehicle may get it to a safe location prior to more damage being inflected on themselves or others.

Keywords: Controller Area Network, Security, Analysis, Monitoring, Automatic Socket Closure

1. Introduction

The present incarnations of vehicular CAN Bus is nothing short of a security (and safety) catastrophe (Sakhuja et al., 2021). There exists designers, developers, and engineers taking great pains to ensure the safety of personnel in nearly every endeavor thought possible. This could be as simple as walking across the street to putting human beings into the hostilities of space. Even in the most dangerous situations, such as those faced in military operations, safety reigns supreme (US DoD, 2014). Through most of human history, this concept of safety has been followed in one form or another (sometimes begrudgingly). However, up until recently, these safety concepts were only limited to tangible effects on someone. Tangible as in have a physical presence that can be touched or felt directly. Now, you no longer need a tangible interface to affect safety. It is possible to affect someone or something from distances never thought possible over thirty years ago without physical constructs.

All critical systems connect through the CAN Bus. It is chosen due to having robust availability, high accuracy, and near-immunity to data corruption caused by a vehicle's electrically noisy environment. However, it is extremely difficult to secure as there are several inherent limitations. One of these is that the available computational power of the attached Electrical Control Units (ECUs) is limited. Another is the requirement for near-real-time reactions to inputs allows inadequate time to scrutinize traffic. All units connected to the bus must assume that each system-level command is authentic. The significance caused by this void in vehicle security is apparent once a person realizes that all vehicle controls depend on a secure CAN Bus. This is similar to the human body's central nervous system. If this communication network is compromised, a person will not be able to control their own vehicles or may even lose control caused by an unknown third-party. This devolves into an uncontrolled slug of metal and wire speeding along our roadways haphazardly.

The application of security measures without violating these system requirements is daunting. Until technology reaches a point where it is economically feasible to apply normal solutions, creativity must be used to find ways to reduce incoming threats. One way is to eliminate attack vectors. By removing the remote vectors from wireless sources, security remits to only the physical vectors. This can be done by sensing the CAN Bus for attack anomalies and then determining if the incident originated from a wireless source. In doing so, it may then remove the threat by halting that signal from being received and allowing the system to inherently recover.

Our approach in this research requires a conglomerate Intrusion Detection System (IDS) apparatus with the ability to perform three actions. First, it will determine if the CAN Bus was attacked. When this is detected, the

IDS will sense what the last grouping of signals were that was received from the wireless interface. It will then p determine if the attacking signal was received via wireless. Next, if it does detect an attacking wireless signal, the fingerprint or ID of that wireless signal will be blocked from reception. Lastly, an indication of the attack will be sent to the vehicle operator via a light, a sound, a haptic, or a multiple of the three. It is imperative the alert occur *before substantial* damage occurs to the vehicle or renders the operation of the vehicle beyond safe recovery. The haptic could be a vibration into the steering wheel or a section of the headrest. Performing this as a proof-of-concept and demonstrating its scalability is the motive for this research.

The remainder of this paper is organized as follows: The next section will discuss the problem faced in securing vehicles. In section 3, the current state of vehicular CAN Bus is discussed and the attack vectors present. Section 4 covers the current state of Intrusion Detection Systems as related to CAN Bus. Section 5 presents the concept of the detect-detect-react methodology proposed in this paper and how it processes potential attacks. The impact discussion, conclusion, and path forward are presented in section 6.

2. Inherent Security Issues with CAN Bus

Even though great emphasis is placed on the near-real time communication network, it is hampered by the ease in which it is to degrade the availability or negatively modify communication integrity. Typical networks have time to detect and sufficient computational power to manage the situation before an attack on them is fatal. There is a significant lack of both in the CAN Bus protocol and set-up (required real-time communication and paltry computer power). The main hinderance to treating CAN Bus communications as a normal network communications is the near-real-time requirement (Wu et al., 2020). There is not sufficient time-overhead to allow for adequate security checks. In addition, current vehicle configurations from OEM's oblige us to install security equipment after the fact. While manufacturers have decidedly put forth safety as their number one consideration, they have fallen short in an extreme way. The main reason is the rabid insecurity that grips the CAN Bus. As stated before, safety has been designed, engineered, and litigated into each vehicle allowed on our roadway's sans the networks. *However, without proper security for the vehicle systems, all the safety mitigations taken could be in vain.* Without proper security, it could bring the main method of personal transportation to a standstill. With modern over-the-air updates, it is possible to convert a fleet of autonomous vehicles into a vehicle equivalent of a botnet (Bagloee et al., 2016).

Now, over thirty years later from when the CAN Bus made its debut, there are over 275 million vehicles registered in the United States (US DoT, 2019, States, 2022). The average age of those vehicles has surpassed 11.5 years (Culver, 2020, Bomey, 2019, Carlier, 2021, Deerwester, 2022). Assuming normal statistical distribution, that would place vehicles on the older part of the curve being made in 1997 and later. This equates to over 200 million vehicles present on U.S. roadways that have unsecured networks.

The protocol used on the CAN Bus does not hold well to the tenants of security. When measuring this protocol against the appeal of the CIA model, it falls short. These are the three essential security criteria. *Confidentially* entails providing information only to those people or entities that are required or cleared to access it. The CAN protocol does not include any mechanism or cryptographic ability to prevent unauthorized persons from accessing all information contained on the bus. *Integrity* deals with the exactitude, legitimacy, and comprehensiveness of the data. The CAN protocol does contain error checking mechanisms at a rudimentary level. However, these methods cannot detect the injection of false data onto the bus. Without that ability, the data on the bus will never attain full integrity. *Availability* involves the system always being accessible by authorized users. With the presence of priority-based messaging system, if a message with a higher or the highest priority is inserted onto the bus, lower priority messages will be barred access to the network. Incidentally, availability is one of the primary strengths of the CAN Bus, but it is easily turned into one of its weaknesses.

3. Controller Area Networks

3.1 Computer Solution for Vehicles

The normal progression of computing led to reducing the physical size of computational devices. It was only a matter of time before microprocessor size became trivial enough to find themselves embedded into everything, including automobiles. When Karl Benz invented what would pass as the first car in 1886, it was little more than a frame, an internal combustion engine, and three wheels (Jardine, 2022). At the turn of the 20th century, cars were still a simple mix of mechanical parts put together to get someone from point A to point B faster than a

horse drawn carriage. As time wore on, they became increasingly more complex. In the sixty years between 1911 and 1971, cars saw the implementation of electronic ignition, air conditioning, power steering, and anti-lock brakes. The additional features were enhanced using electronics and more wiring was needed. This led to increased weight. This affected fuel mileage which became an issue during the 1973-1974 oil embargo. With the federal government prompting cars have better fuel mileage, a solution was needed (Day, 2016).

3.2 CAN Bus is Created:

Robert Bosch GmbH created the CAN Bus in 1985 providing a solution for electronic devices to communicate without the requirement of separate lines (Corrigan, 2016) (Automation, 2018). The result allowed components to communicate without needing dedicated physical wires between them. The components could share one medium. This eliminated the need for additional wires for each communication line and lowered cost and weight. In 1991, Mercedes Benz was the first to utilize this framework in the mass production of vehicles (the W140) (Daimler, 2016).

The CAN Bus is an asynchronous serial communications network. The physical characteristics of the CAN Bus is exceedingly unassuming. This is one of the motivations for its appeal. There must be two or more Electronic Control Units (or ECU's) on the bus for the protocol to function correctly. Each node is connected to the conventional two-wire, twisted pair bus. Both wires have a potential of 2.50V that flow in opposite directions to each other. Along with the twisted pair, the current flowing in opposite directions ensure that electronic noise effects are minimized. This configuration provides near immunity to the electronic noise that is prevalent in a car's environment. There is not a central hub or routing system, just a continuous flow of information that's always available (Wojdyla, 2012).

3.3 Attacks and Vulnerabilities

Contemporary intrusions and vehicular attacks have been physical attacks that require hands-on access to the vehicle (cutting a brake line or hotwiring a car) (Nilsson, 2008). These types of infiltrations normally leave some residual evidence that, upon inspection, will reveal that a malicious incident has been exercised or has been attempted. It is entirely different when it comes to attacks on the car's CAN Bus. Physical manifestations of these types of attacks are not as clear. Since these attacks do not rely on physical contact, they can be executed and leave little to no evidence. These attacks would be similar to SYN-Flood attacks on a Wi-Fi network (Lutkevich, 2022). Attack detection is possible. However, it takes more effort and ability than what is required for a simple visual inspection of the attack site. The inspector would need full understanding of how the attack took place and what to look for in reviewing logs. In a sense, without cause to think what had happened, these attacks could occur without leaving a noticeable trace.

Attack vectors on the CAN Bus have two overarching classifications. One is physical and the other is not (wireless). Physical attacks require access to the vehicle. Wireless attacks do not. Securing the wireless portion would result in the possible removal of that vector. That would essentially return cybersecurity to that of a physical issue as it was prior to the proliferation of installed wireless technology in vehicles. Therefore, the focal point is decreasing the possibility of success through the CAN Bus wireless attack vector (Nilsson et al., 2008).

According to (Wu et al., 2020), CAN Bus attacks generally have three steps. First step is to target the vehicle. This can be done either physically or wirelessly. The second step is to compromise the CAN Bus itself or a device on the CAN Bus. This may be performed by flashing firmware on an ECU or by transmitting bogus frames. The final step is what the attacker wants to do with the control. Do they want to start/stop/damage the vehicle, steal information, or just to wreak havoc? It is vital when researching the attack potential of the CAN bus, as well as individual ECUs, that all pathways into and out of the system in question be evaluated. The most critical attacks are replay, injection, and/or denial of service (Batmaz and Pete, 2018). The assumption is made that an attacker is not limited to the computational power of an ECU but can use more powerful devices for attacks (Luo and Hou, 2019b).

The aggregate of attack research performed on the core of the CAN Bus is extensive. The notion of placing an Intrusion Detection System (IDS) in this communication network is not a novel approach. Table 1 below outlines several attacks performed during past CAN Bus security research. Please note, two-thirds of the attacks do not require access to the vehicle.

Table 1: Attacks Performed on CAN Bus

Research	Interface	Physical Access	Method
(Zhang et al., 2016)	OBD-II	Yes	ELM327 OBD-II Bluetooth device and malicious application on a smartphone
(Kang et al., 2018)	OBD-II	Yes	Arduino UNO with CAN bus connects to OBD-II port
(Studnia et al., 2013)	USB	Yes	Connect to media systems by using smartphone
(Oka et al., 2014)	Bluetooth	No	Using the PIN vulnerability while paring
(Josephial and Adepu, 2019)	Wi-Fi	No	Scanning vulnerabilities through appropriate tools such as Nmap, Nessus, and Metasploit
(Woo et al., 2015)	3G/4G	No	Transmit controlling frames by smartphone connecting to CAN via 3G/4G network
(van der Heujden et al., 2017)	V2X	No	DoS at data injection against VANET
(Chattopadhyay et al., 2018)	V2X	No	Injection and packet drop attacks
(Petit et al., 2015)	Sensors	No	Attacking LiDAR, radar, and camera sensors remotely based on physical vulnerabilities

4. Intrusion Detection Systems (IDS)

Overall, the current literature discusses three overall possibilities to provide CAN Bus security. First is the use of cryptography. It has been discussed as a possible way of ensuring frame integrity and confidentiality. Another is the insertion of a firewall between critical and non-critical sections of the CAN Bus. Third is adding an IDS at the entry points of the CAN Bus (Wu et al., 2020). With our current technology, application of the first two increase the required response times to an unacceptable level. It is in the third extent that this paper dwells to find a possible solution.

To adequately detect an event on the CAN Bus, three requirements must be met. First, there must be a device to detect the event. This can be one of the detection mechanisms already presented in the literature or a combination of a few of them. Second, there must be a device that can store the information found during the detected event. This could be volatile or non-volatile. In addition, the amount of needed storage will fluxuate depending on how much past information is desired to be kept. Finally, there must be a device that can interpret the bus traffic for storage on the device. This has been referred to as a “decision engine” in the current intrusion detection literature. These requirements may be met with a single device or combination of devices (Moustafa et al., 2019).

Figure 1 (Luo and Hou, 2019b), provides a basic diagram of the vehicle network and how it interfaces with external sources. There are three distinct layers or areas to note. The first is the External Wireless Interface layer which comprises of the transmission and reception of all the wireless signals. The demarcation for this layer is just inside the Wireless Gateway/Central Gateway. It encompasses the wireless signals as well as the physical connections. The second layer, called the Internal Vehicle Network (IVN) layer by the authors, contains all the network systems. This incorporates not only the target CAN Bus but also the other various networks. The third is External Physical Interface layer. This involves the connections to the physical items that an operator or a person can physically attach to the IVN layer. Potential cyberattacks are then analyzed at each layer and corresponding countermeasures are presented.

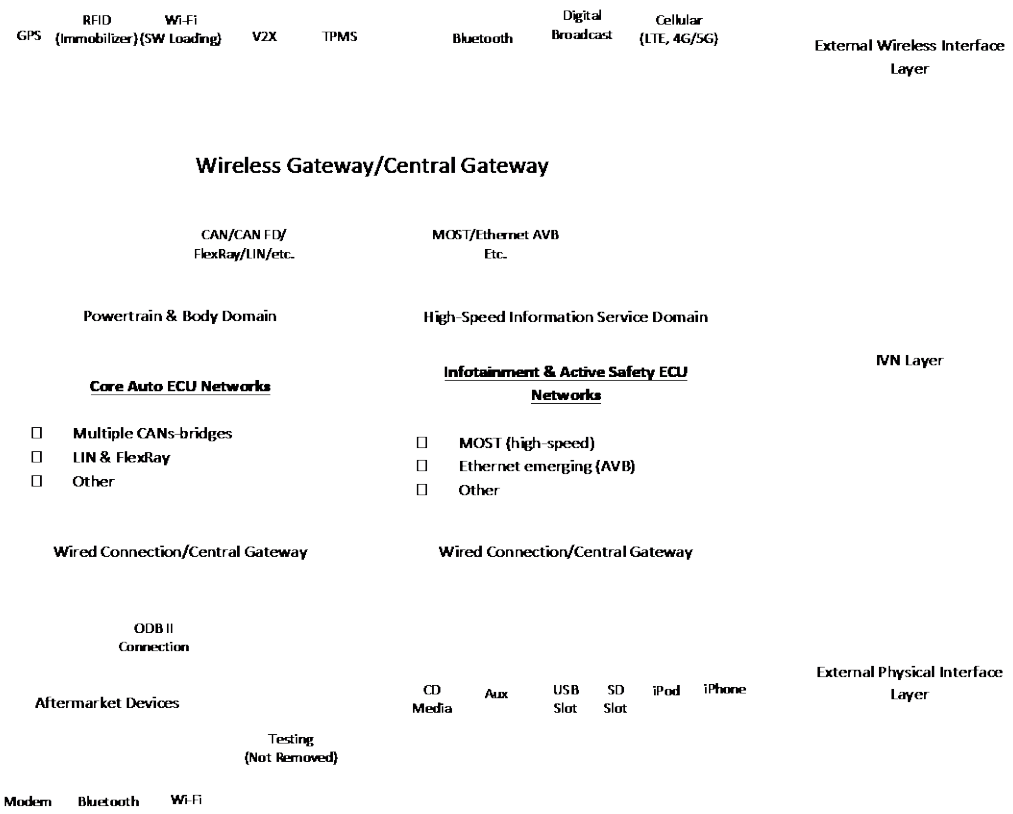


Figure 1: Three-layer structure of the IVN structure from the perspective of external interface (Luo and Hou, 2019a)

Based on current IDS research, there are a few types of sensors that appear to be feasible for CAN Bus (Luo and Hou, 2019b). They are broken down into:

- Dataflow Based: these check only the identification values of CAN frames and thus are more suitable for real-time detection. They can identify cyberattacks but require substantial computing power. Examples include the frequency of communications, ECU fingerprinting, and systematic entropy.
- Payload Based: these check the actual payload of each frame. These methods rely on the utilization neural networks and other forms of machine learning. However, it is not known how well these types would perform in the automotive environment. Examples of these include frame ID checks and checksum verifications.
- Hybrid: methods that extract the best of both worlds in dataflow and payload-based types. These require specialized hardware and are a prime utilization for FPGA's (Luo and Hou, 2019b). These may also employ a total of eight specialized sensors as notated in Table 2.

Table 2: Intrusion Detection Sensors

Sensors	Description
Formality	Correct message size, header and field size, field delimiters, checksum, etc.
Location	Message is allowed with respect to a dedicated bus system
Range	Compliance of payload in terms of data range
Frequency	Timing behavior of messages is approved
Correlation	Correlations of messages on different bus systems adheres to specification

Sensors	Description
Protocol	Correct order, start time, etc. of internal challenge-response protocol
Plausibility	Content of message payload is plausible, no infeasible correlation with previous values
Consistency	Data from redundant sources is consistent

5. Concept of Providing Security

This research focuses on the wireless connectivity of the vehicle. As shown in Figure 1, there are a considerable number of external communication vectors. Intrusion detection occurs on one of the CAN Buses and results in the potential blocking of a communication vector. The perfect answer to the security drawback would be to have confidentially, integrity, and availability (CIA) preserved as is done in standard networking. However, the technology and knowledge in doing that successfully, while overcoming the limitations and restrictions of the system, is currently unavailable. Nevertheless, if it is possible to detect and alert the operator of an intrusion at this gateway, all wireless vectors would be protected.

To provide a graphic explanation of this research, figure 2 shows a functional layout and shows where the sensing and the feedback loops would be connected. In addition, it shows where each element would be located, and demonstrates how the apparatus would be applied. Overall, the concept will continuously monitor the CAN Bus for pre-set anomalies, like how antivirus software utilizes a signature database. If an anomaly is detected, the apparatus will then review the incoming wireless signals to determine if that was the source of the anomalous injection. If it is the source, steps could be taken to pinpoint the port/socket and close it.

5.1 Operational Concept

Figure 2 is a diagram of the proposed research apparatus. It has been adopted from (Luo and Hou, 2019a) and manipulated to suit the needs of this research. There are three main sections. Along the top is the wireless gateway where the universe of RF signals interface with the vehicle. Following from this on the left side are the vehicle's CAN bus that connects the various ECU's. Both items are found already installed in today's vehicles. It is the block diagram to the right where the research is to take place. This block is broken down into various sub-functions which outline what happens. Here, the device will monitor and log the wireless traffic into and out of the vehicle. In addition, the CAN Bus is sensed for any anomalies present. With the current iteration, the checks it does are akin to how a virus scanner performs their inspections. It searches for potential known signatures of attacks.

At the outset, the incoming wireless signals are parsed out and logged as they are received at the Wireless Gateway. This provides a history of what has been received. It will be used for review in the event an anomaly is detected on the CAN Bus. From there, the CAN Bus is monitored for any anomalous signals. The current iteration of the research contains simple triggers based on what has been observed in the past on that network. This is akin to a signature type of detection. This appears to be the easiest method to ensure the concept works. The CAN Bus is being used as the area for anomaly detection rather than the raw incoming wireless signals. This is due to the relative ease of detecting an abnormality with respect to the incoming wireless signals. CAN Bus communications are predictable and mundane as compared to Wi-Fi, Bluetooth, and cellular signals. In the event a potential anomaly is detected, a decision engine is enacted to see if what has been encountered can be classified as an anomaly. This is done by the detected potential anomaly being compared to a signature of known attacks. Previous research such as that from (Dwivedi, 2022), (Han et al., 2018), and (Marchetti and Stabili, 2017) demonstrate such signatures exist for the CAN Bus anomalies.

Once an attack has been established, the determination is made as to the attack origination. It can either be from an external source or from an embedded internal source. This is done by evaluating the incoming wireless signals that were logged earlier. If a received signal adequately maps to a potential structure that could cause the anomaly, a command is sent to the Wireless Gateway to initiate a closure of that communication socket. For example, imagine there exists an interpreted signal on the CAN Bus that directs the steering wheel to instantly turn 90 degrees while the car is going 80 MPH, that could be determined to be an abnormal signal. This would be especially true if that signal kept repeating and not deviating. The detection system would review the logged

incoming signals to determine if an inbound wireless signal is directing that resulting frame on the CAN Bus. If that is the case, a command is sent to the Wireless Gateway to suspend that communication socket. In addition, the apparatus will alert the operator that an abnormal wireless signal was received and caused irregular indications on the CAN Bus. This can be a simple indicator light on the instrument cluster, an alarm that will interrupt and play through the speakers (maybe a voice), or a haptic all depending on the desires of the operator.

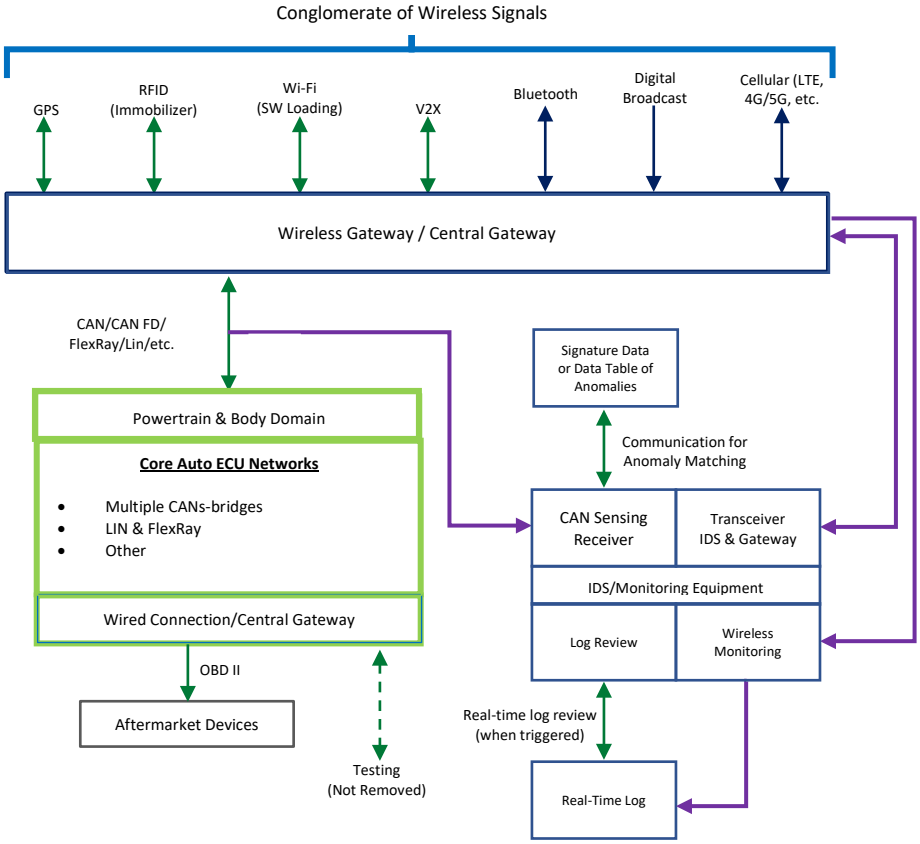


Figure 2: Projected Layout & Function of Research Apparatus

6. Way Forward and Conclusion

The concepts presented in this paper are not new. The act of monitoring and analyzing both CAN Bus as well as wireless signals have been performed exhaustively in the past. The paring of these methods together has not been well documented and researched. As a follow-on to this research, our future work entails the development of a test bed to run experimental evaluation on the concept. When this is successful, the addition of successive actions should be relatively easy. Once able to alert the driver, a wireless cutoff could be installed. These are utilized in Virtual Private Networks (VPN) today when it loses connectivity. This prevents the network protected by the VPN to become active without the VPN in place.

References

AUTOMATION, C. I. 2018. *History of CAN Technology* [Online]. CAN in Automation. Available: https://www.can-cia.org/can-knowledge/can/can-history/ [Accessed 15 April 2022].

BAGLOEE, S. A., TAVANA, M., ASADI, M. & OLIVER, T. 2016. Autonomous Vehicles: Challenges, Opportunities, and Future Implications for Transportation Policies. *journal of Modern Transportation,* 24**,** 284 - 303.

BATMAZ, G. & PETE, I. 2018. Controller Area Network (CAN) Deep Packet Inspection. Nvidia.

BOMEY, N. 2019. Old Cars Everywhere: Average Vehicle Age Hits All-Time High. *USA Today*, 28 June 2019.

CARLIER, M. 2021. Age of U.S. Light Vehicles 2002-2020. *In:* TRANSPORTATION, D. O. (ed.). Statista.com: US Department of Transportation.

CHATTOPADHYAY, A., MITRA, U. & STROM, E. C. 2018. Secure Estimation in V2X Networks with Injection and Packet Drop Attacks. *15th International Symposium on Wireless Commuication Systems (ISWCS).* Lisbon, Portugal: IEEE.

CORRIGAN, S. 2016. Introduction to the Controller Area Network (CAN). Texas Instruments.

CULVER, M. 2020. Average Age of Cars and Light Trucks in the U.S. Approaches 12 Years, According to IHS Markit. IHS Markit.

DAIMLER, C. N. M. 2016. *Mercedes W140: First Car with CAN* [Online]. CAN Newsletter Magazine. Available: https://can-newsletter.org/engineering/applications/160322_25th-anniversary-mercedes-w140-first-car-with-can [Accessed 10 March 2022].

DAY, C. 2016. Computers in Cars. *Computing in Science and Engineering.* IEEE CS American Institute of Physics.

DEERWESTER, J. 2022. The Average Age of a Car in the US is up to 12.2 Years, a New Record. How Old is Yours? *USA Today*, 24 May 2022.

DWIVEDI, A. K. 2022. Anomaly Detection in Intra-Vehicle Networks. arXiv.

HAN, M. L., KWAK, B. I. & KIM, H. K. 2018. Anomaly Intrusion Detection Method for Vehicular Networks Based on Survival Analysis. *Vehicular Communications,* 14**,** 52-63.

JARDINE. 2022. *The History of Car Technology* [Online]. Jardine Motors Group. Available: https://news.jardinemotors.co.uk/lifestyle/the-history-of-car-technology [Accessed 08 March 2022].

JOSEPHIAL, E. F. M. & ADEPU, S. 2019. Vulnerability Analysis of an Automotive Infotainment System's Wifi Capability. *19th International Symposium on High Assurance Systems Engineering (HASE).* Hangzhou, China: IEEE.

KANG, T. U., SONG, H. M., JEONG, S. H. & KIM, H. K. 2018. Automated Reverse Engineering and Attack for Can Using OBD-II. *IEEE 88th Vehicular Technology Conference (VTC-Fall).* Chicago, IL, USA: IEEE.

LUO, F. & HOU, S. 2019a. Cyberattacks and Countermeasures for Intelligent and Connected Vehicles. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems,* 12**,** 55 — 66.

LUO, F. & HOU, S. 2019b. Cyberattacks and Countermeasures for Intelligent and Connected Veicles. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems,* 12**,** 55 — 66.

LUTKEVICH, B. 2022. *SYN Flood Attack* [Online]. TechTarget. Available: https://www.techtarget.com/searchsecurity/definition/SYN-flooding [Accessed 15 January 2023].

MARCHETTI, M. & STABILI, D. 2017. Anomaly Detection of CAN Bus Messages Through Analysis of ID Sequences. *IEEE Intelligent Vehicles Symposium (IV).* Los Angeles, CA, USA: IEEE.

MOUSTAFA, N., HU, J. & SLAY, J. 2019. A Holistic Review of Network Anomaly Detection Systems: A Comprehensive Survey. *Journal of Network and Computer Applications,* 128**,** 33 - 55.

NILSSON, D. K. 2008. Secure Firmware Updates of the Air in Intelligent Vehicles. *IEEE International Conference on Communications Workshops.* Beijing, China: IEEE.

NILSSON, D. K., LARSON, U. E. & JONSSON, E. 2008. Efficient In-Vehicle Delayed Data Authentication Based on Compound Message Authentication Codes. *!EEE 68th Vehicular Technology Conference.* Calgary, BC, Canada: IEEE.

OKA, D.-K., FURUE, T., LANGENHOP, L. & NISHIMURA, T. 2014. Survey of Vehicle IoT Bluetooth Devices. *7th International Conference on Service-Oriented Computing and Applications.* Matsue, Japan: IEEE.

PETIT, J., STOTTELAAR, B., FEIRI, M. & KARGL, F. 2015. Remote Attacks on Automated Vehicles Sensors: Experiments on Camera and Lidar. *Black Hat Europe.* Amsterdam, The Netherlands: Security Innovation.

SAKHUJA, S., DUNNE, M. & FISCHMEISTER, S. 2021. The Boy Who Cried Wolf: On Precision in CAN Bus Instrusion Detection. *Embedded Security in Cars - USA.* Virtual.

STATES, U. 2022. Number of Motor Vehicles Registered in the United States from 1990 to 2020. *In:* TRANSPORTATION, D. O. (ed.). Federal Highway Administration.

STUDNIA, I., NICOMETTE, V., ALATA, E., DESWARTE, Y., KAANICHE, M. & LAAROUCHI, Y. 2013. Survey on Security Threats and Protection Mechanisms in Embedded Automotive Networks. *43rd Annual IEEE/IFIP Conference on Dependable Systems and Networks Workshop (DSN-W).* Budapest, Hungary: IEEE.

US DOD, D. O. D. 2014. DoD Safety and Occupational Health (SOH) Program. *In:* DEFENSE (ed.). Washington, D.C., USA: U.S. Government.

US DOT, D. O. T. 2019. Highway Statistics 2018. *In:* U.S. DOT, D. O. T. (ed.). Washington, D.C.: Federal Highway Administration.

VAN DER HEUJDEN, R., LUKASEDER, T. & KARGL, F. 2017. Analyzing Attacks on Cooperative Adaptive Cruise Control (CACC). *IEEE Vehicular Networking Conference (VNC).* Turin, Italy: IEEE.

WOJDYLA, B. 2012. *How it Works: The Computer Inside Your Car* [Online]. Hearst Digital Media: Popular Mechanics. Available: https://www.popularmechanics.com/cars/how-to/a7386/how-it-works-the-computer-inside-your-car/# [Accessed 02 February 2021].

WOO, S., JO, H. J. & LEE, D. H. 2015. A Practical Wireless Attack on the Connected Car and Security Protocol for In-Vehicle CAN. *IEEE Transactions on Intelligent Transportation Systems,* 16**,** 993 - 1006.

WU, W., LI, R., XIE, G., AN, J., BAI, Y., ZHOU, J. & LI, K. 2020. A Survey of Intrusion Detection for In-Vehicle Networks. *IEEE Transactions on Intelligent Transportation Systems,* 21**,** 919 — 933.

ZHANG, Y., GE, B., LI, X., SHI, B. & LI, B. 2016. Controlling a Car Through OBD Injection. *IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud).* Beijing, China: IEEE.

Digital Insanity: Exploring the Flexibility of NIST Digital Identity Assurance Levels

Kenneth Myers
School of Technology and Innovation, Marymount University, Arlington, Virginia, USA

Kmm57090@marymount.edu

Abstract: NIST Special publication 800-63-3 present a new risk management concept on digital identity. It includes various harm categories to determine an appropriate assurance level for identity proofing, authentication, and federation. These three distinct approaches are highlighted to give flexibility in protecting systems. This paper explores if this is a realized flexibility through developing a tool to test assurance level and component flexibility. It also identifies appropriate MFA levels given different levels of risks and makes three recommendations to help improve adoption of the NIST digital identity guidelines.

Keywords: Assurance Level, Digital Identity Risk Assessment, Digital Identity, Identity Proofing, Authenticator, Federation

1. Introduction

Digital identity risk management is a new field of risk management. It ensures that the right person has access at the right time and supplements existing NIST Risk Management Framework processes. NIST first introduced Digital identity risk management through the third revision of Special Publication 800-63 in 2017. Similar to other risk management frameworks, it identifies specific controls based on an assessed adverse impact level. This identity risk management framework is unique because it dissects digital identity assurance into three elements: identity, authenticator, and federation assurance levels.

1. Identity assurance - identity proofing process.
2. Authenticator assurance - authentication process.
3. Federation assurance - assertion in a federated environment.

The reason for the dissection is that "combining proofing, authenticator, and federation requirements into a single bundle sometimes have unintended consequences and can put unnecessary implementation burden on the implementing organization" (Grassi, Garcia and Fenton, 2017, p.17). Another benefit is the flexibility in using higher assurance authentication without requiring high assurance identity proofing based on the risk assessment.

- Research question 1 - Does dissecting digital identity into specific components make the component risk assessment more flexible?
- Research Questions 2- Should all transactions use an MFA?
- Research Question 3 - Should all transactions use phishing-resistant MFA?

This dissection approach in identity control families is like other risk management frameworks.

- NIST Risk Management Framework control family.
- Cyber Security Framework subcategory.
- Center for Internet Security Critical Security Controls control group.

Within the NIST digital identity guidelines is a digital identity risk management outlining impacts per harm category. There are six harm categories used to determine an appropriate assurance level.

1. Potential for inconvenience, distress, or damage to standing or reputation.
2. Potential impact on financial loss.
3. The potential impact of harm to agency programs or public interest.
4. The potential impact of unauthorized sensitive information release.
5. Potential impact on personal safety.
6. The potential impact of civil or criminal violations.

Who should conduct this risk assessment, such as the application owner, a risk professional, or a cross-functional team? A user must select a risk level from low to high within each harm category. Low means limited, short-term, while high means severe long-term impact. NIST includes a flow chart to help determine a level and shortcuts, but the process could be more intuitive.

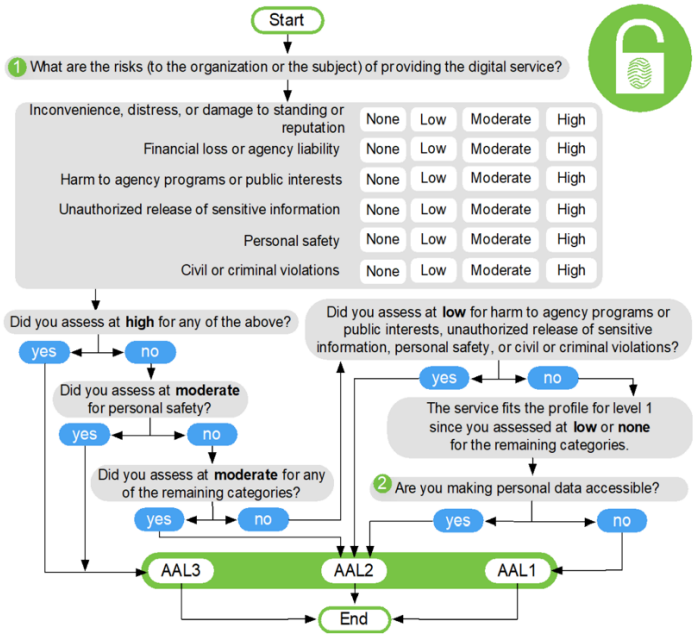


Figure 1: NIST 800-63-3 Authenticator Assurance Level Flow Chart (Grassi, Garcia and Fenton, 2017)

In 2020, the Identity, Credential, and Access Management Subcommittee formed a digital identity risk assessment working group to develop an identity risk assessment playbook. This playbook (GSA, 2021) outlines a process to conduct a risk assessment and record the findings in a risk acceptance statement. The playbook also contains more intuitive flow charts.

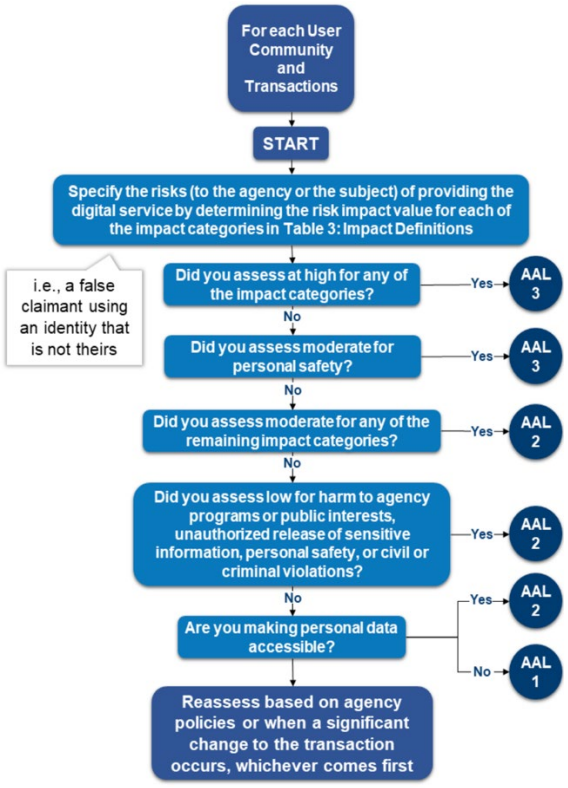


Figure 2: GSA Digital Identity Risk Assessment Playbook Authenticator Assurance Level Flow Chart (GSA, 2021)

The following section highlights adoption trends identified through a literature review.

2. Literature Review

Adopting the NIST Digital Identity Risk Management guidelines appears limited both within the Federal Government and outside. Federal agencies must implement Digital Identity Risk Management and any successive versions according to Office of Management and Budget Memo 19-17 (2019). One may assume agencies have fully adopted these NIST identity guidelines. However, an OMB memo does not have the same authority as a public law like the Federal Information Security Management Act (FISMA). FISMA requires agencies to implement the NIST Risk Management Framework to authorize IT systems (S.2521, 2014). Despite this, there is some indication of government-wide adoption similar to the NIST Risk Management Framework or the Federal Risk and Authorization Management Program. A few leading indicators of government and industry use point to this assessment.

- Other NIST risk management frameworks have multiple agency or industry adoption examples. Nine other languages have translated the NIST Cybersecurity Framework (2022)! The Digital Identity Framework has limited informative resources but does include a frequently asked question sheet, implementation resources, and conformance criteria.
- OMB memo 19-17 points to idmanagement.gov for a list of trusted services that align with NIST 800-63-3. The trust services page on idmanagement.gov only lists identity credentialing services (GSA, 2022a). There needs to be a list of services at various assurance levels.
- NIST requires an MFA when personal information is stored and when self-asserted (Grassi, Garcia and Fenton, 2017, p. 31). A quick scan of publicly available federal login sites that require a user's personal information to create an account, such as recreation.gov, some sba.gov login sites (Capital Access Financial System and Shuttered Venue Operators Grant, and Restaurant Revitalization Award Portal), and the federal training website fedvte.usalearning.gov allow someone to create an account or login without using an MFA. There is also a large number that does enforce MFA through integration with login.gov.

Some organizations have published supplemental material as well.

- MITRE (2020) published a methodology for unsupervised remote identity proofing (identity assurance level 2) and supervised remote in-person proofing (identity assurance level 3).
- NSA (2020) published a fact sheet that assessed multi-factor authenticator solutions to an Authenticator Assurance Level.
- The DoD ICAM Reference Design (2020) includes digital identity assurance levels.

Some vendors also have supplemental material on how their products align with the NIST guidelines.

- Microsoft (2022a) published a detailed approach to implementing authenticator assurance levels.
- Forgerock (2021) published a white paper on their product alignment with identity, authenticator, and federal assurance levels.

One common trend in almost all materials is a need for more clarity in assessing an assurance level. In other words, all material helps explain each level and conformance criteria on each level, but little to help explain how to determine a level. This is necessary because multiple examples of companies choosing the wrong authenticator type exist. Not all MFA are secured equally.

A very prevalent attack today is an MFA bypass or fatigue attack. This technique uses repeated push notifications to fatigue the user into accepting the request (Cimpanu, 2022). This new attack is used throughout 2021. Another typical authenticator attack is an MFA interception attack. This attack style intercepts an MFA piece of knowledge, such as an SMS pin or mobile application one-time PIN (Jansen, 2021), for an attacker to replay. An MFA interception is sometimes combined with a SIM jacking attack so that an SMS pin will arrive on an adversary's device. These types of authenticators are Authenticator Assurance Level 2.

Microsoft highlights specific accounts that are attractive for credential theft (2022b). These accounts include the following.

1. Administrator accounts with escalated security privileges, such as domain accounts or help desk personnel.
2. Manager accounts to approve access requests or other higher-risk transactions.
3. Very Important People include executives, legal staff, product planners, and researchers.

These higher-risk accounts roughly align with the same account types identified in the U.S. Government Privileged Identity Playbook (2022). These risky accounts should have the strongest authenticator or Authenticator Assurance Level 3. What about general user accounts? As mentioned earlier, following the NIST risk criteria, the risk of a general user account is most likely an Authenticator Assurance Level 2, which is susceptible to phishing. The following section outlines the research approach to explore the research question.

3. Research Method

One identified drawback of the NIST digital identity guidelines is the complicated nature of conducting an assessment. This is consistent in both the NIST guidelines, but also in the Digital Identity Risk Assessment Playbook. This paper developed a simple, excel-based tool to make the process more user-friendly. The tool's purpose is to test and verify the flexibility of the three identity components. This tool has the following features.

1. A consolidated assessment interface that combines identity, authentication, and federation assurance levels.
2. A user adjusts a set of drop-down menus to pick the desired risk level.
3. A calculation to automatically determine the correct assurance level based on the NIST digital identity guidelines flow charts.
4. A terminology tab of harm categories and risk descriptions.

Test 5 - PII Yes, Personnal safety moderate				2	1	1
Is PII or PHI collected and does it need to be validated?	*Yes & I Don't Know*			2		
Impact Category	**Select Impact Definition***	**Impact Level**		**IAL**	**AAL**	**FAL**
Inconvenience, distress, or damage to standing or reputation to the agency	*t worst, limited, short-term inconvenience, distress, or embarrassment to any part*	Low		1	1	1
Financial loss or agency liability	*r inconsequential financial loss to any party, or at worst, an insignificant or incons*	Low		1	1	1
Harm to agency programs or public interests	*Not Applicable*	Not Applicable		1	1	1
Unauthorized release of sensitive information	*Not Applicable*	Not Applicable		1	1	1
Personal safety to users	*Not Applicable*	Not Applicable		1	1	1
Civil or criminal violations	*Not Applicable*	Not Applicable		1	1	1

Figure 3: NIST 800-63-3 Digital Identity Risk Assessment Tool

A series of test conditions were used in the tool to test the flexibility of both the components and the assurance levels.

- Component flexibility is the potential to test each component independently.
- Assurance level flexibility is the potential to test each assurance level's independence from the other.

Six test cases were developed and run through the tool. Only two variables are tested on whether the transaction included Personally Identifiable Information (PII) and one randomly chosen harm category and level.

1. Includes PII, All categories low or N/A
2. No PII, Any category moderate except personal safety.
3. Includes PII, Sensitive information high
4. Includes PII, Financial Loss high
5. Includes PII, Personal safety moderate
6. No PII, All categories low or N/A

The following section provides the test results and analysis.

4. Analysis

The following results were determined using the tool.

Table 1: Risk Assessment Results

Test Case	Identity Assurance Level	Authenticator Assurance Level	Federation Assurance Level
#1. Includes PII, All categories low or N/A	2	2	1
#2. No PII, Any category moderate except personal safety.	2	2	2
#3. Includes PII, Sensitive information high	3	3	3
#4. Includes PII, Financial Loss high	3	3	3
#5. Includes PII, Personal safety moderate	3	3	3
#6. No PII, All categories low or N/A	1	1	1

Two interesting observations from the tests.

1. There is component flexibility. They can do this if someone wants to assess only one or two of the three components.
2. There does not appear to be assurance-level flexibility. If an assessed harm category is high or moderate, this level persists through the identity components.

To answer research question 1, dissecting digital identity into specific components does give component flexibility but not level flexibility. The two remaining research questions are answered through the literature review and the test results. The only test cases where MFA (Authenticator Assurance Level 2) was not required were test cases where no personal data and no risk in the transaction. It is a recommendation from the Digital Identity Risk Assessment Playbook not to consider a login based on these indicators (GSA, 2021). The question of whether all accounts should use phishing-resistant MFA points to yes. Most current forms of MFA, including various one-time pins, are susceptible to MFA interception and fatigue-style attacks. A phishing-resistant authenticator disrupts adversary in the middle tactics to steal credentials.

This research paper arrived at three additional recommendations.

1. If the federal government wants wide adoption, consider including the NIST digital identity guidelines in public law or convert it to a FIPS publication. Either approach would require federal agencies to implement digital identity risk management.
2. Given the susceptibility of most Authenticator Assurance Level 2 authenticators, NIST should update their guidelines only to specify phishable MFA options for non-enterprise use cases. In contrast, phishing-resistant MFA should be used for all enterprise or low to moderate-risk transactions.
3. Since the harm categories are consistent across each assurance level, consolidate the risk determination flow charts.

5. Conclusion

This research explored the risk management subfield of digital identity. It tested the flexibility of the component approach but also the assurance levels. Six test cases were developed and tested and found a need for more flexibility in assurance levels. This paper also explored two additional research questions if all accounts need MFA and if all accounts should only use phishing-resistant MFA. Multiple attack techniques highlight the susceptible nature of one-time pin and push notifications to compromise. Additionally, the NIST risk assessment highlights that all transactions with PII and low to moderate risk should have MFA.

This research makes three additional recommendations.

1. If the federal government wants wide adoption, consider a public law that requires federal agencies to implement digital identity risk management.
2. Given the susceptibility of most Authenticator Assurance Level 2 authenticators, NIST should update their guidelines only to specify phishable MFA options for non-enterprise use cases. In contrast, phishing-resistant MFA should be used for all enterprise or low to moderate-risk transactions.
3. Since the harm categories are consistent across each assurance level, consolidate the risk determination flow charts.

This research can be further extended through continued development of a tool into a public website resource or into a GitHub code repository. Additionally, explore whether two authenticator assurance levels, phishable and phishing—resistant are a more modern approach.

Acknowledgements

I want to thank my cyber risk professor, Dr. Michelle Liu, for mentoring me through the development of this paper.

References

Cimpanu, C. (2021) *Russian hackers bypass 2FA by annoying victims with repeated push notifications*. *The Record by Recorded Future*. Available at: https://therecord.media/russian-hackers-bypass-2fa-by-annoying-victims-with-repeated-push-notifications/ (Accessed: December 2, 2022).

DoD. (2020) *Department of Defense Enterprise ICAM Reference Design*. DoD CIO. Available at: https://dodcio.defense.gov/Portals/0/Documents/Cyber/DoD_Enterprise_ICAM_Reference_Design.pdf (Accessed: December 1, 2022).

Forgerock. (2021) *ForgeRock and NIST Special Publication 800-63-3*. Available at: https://www.forgerock.com/resources/whitepaper/forgerock-and-nist-special-publication-800-63-3 (Accessed: November 20, 2022).

Grassi, P. A., Garcia, M. E., and Fenton, J. L. (2017). *Digital identity Guidelines*. Special Publication, 800(63–3). Available at: https://doi.org/10.6028/nist.sp.800-63-3/ (Accessed: November 20, 2022).

GSA. (2021, November 17). *Digital Identity Risk Assessment Playbook*. FICAM Architecture. Available at: https://playbooks.idmanagement.gov/playbooks/dira/ (Accessed: December 2, 2022).

GSA. (2022a, September 21). *Trust Services*. idmanagement.gov. Available at: https://www.idmanagement.gov/buy/trust-services/ (Accessed: November 22, 2022).

GSA. (2022b, September 25). *Privileged Identity Playbook*. idmanagement.gov. Available at: https://playbooks.idmanagement.gov/playbooks/pam/ (Accessed: November 22, 2022).

Jansen, W . (2021, January 12). Abusing cloud services to fly under the radar. Available at: https://research.nccgroup.com/2021/01/12/abusing-cloud-services-to-fly-under-the-radar/ (Accessed: November 23, 2022).

Microsoft. (2022a, September 15). *Azure Active Directory identity standards overview - Microsoft Entra*. Microsoft Learn. Available at: https://learn.microsoft.com/en-us/azure/active-directory/standards/standards-overview/ (Accessed: November 23, 2022).

Microsoft. (2022b, June 8). *Attractive Accounts for Credential Theft*. Microsoft Learn. Available at: https://learn.microsoft.com/en-us/windows-server/identity/ad-ds/plan/security-best-practices/attractive-accounts-for-credential-theft (Accessed: November 26, 2022).

MITRE. (2020, May 1). *Enrollment and Identity Proofing Practices Statement Templates*. MITRE. Available at: https://www.mitre.org/news-insights/publication/enrollment-and-identity-proofing-practices-statement-templates/ (Accessed: November 25, 2022).

NIST. (2022, November 9). *Cybersecurity Framework Version 1.1*. NIST. Available at: https://www.nist.gov/cyberframework/framework-documents/ (Accessed: November 21, 2022).

NSA. (2020, September 22). *Selecting Secure Multi-factor Authentication Solutions*. Cybersecurity Information. Available at: https://media.defense.gov/2020/Sep/22/2002502665/-1/-1/0/CSI_MULTIFACTOR_AUTHENTICATION_SOLUTIONS_UOO17091520.PDF (Accessed: November 27. 2022)

OMB. (2019, May 21). *Memorandum 19-17 Enabling Mission Delivery through Improved Identity, Credential, and Access Management.* Executive Office of the President. Available at: https://www.whitehouse.gov/wp-content/uploads/2019/05/M-19-17.pdf (Accessed: November 27, 2022).

S.2521 - 113th Congress (2013-2014): Federal Information Security Modernization Act of 2014. (2014). Available At: https://www.congress.gov/bill/113th-congress/senate-bill/2521/ (Accessed: November 23, 2022).

Risk Likelihood of Planned and Unplanned Cyber-Attacks in Small Business Sectors: A Cybersecurity Concern

Tabisa Ncubukezi
Information Technology Department, Faculty of Informatics and Design, Cape Peninsula University of Technology, Cape Town, South Africa

Ncubukezit@cput.ac.za

Abstract: Human factors such as planned and unplanned cyber-attacks are a serious threat to any institution. The presence of planned and unplanned actions exposes the state of cybersecurity within the small business sector – leaving them vulnerable to a range of cyber-risks. This study used AgenaRisk package with Bayesian Network (BN) tools to illustrate the likelihood of risk in planned and unplanned attacks. Adopting the package demonstrates the dependent and independent variables of the human factors, which are planned and unplanned, with their relationships resulting in the ultimate data breach. The work also combined qualitative research with quantitative risk analysis techniques to determine the risk likelihood of the planned activities and unplanned employee actions and their behaviors influencing data breaches. The work used the judgemental sampling method to select twenty-five (25) research participants who are business owners, and Information Technology (IT) managers. An online survey was used to collect data from the selected research participants. Results were analysed using content analysis, and interpreted using the package with BN tools, and risk analysis techniques. The results were further discussed, and the study concluded with the remarks and future developments.

Keywords: AgenaRisk package, Bayesian Network tools, cyber-risks, cybersecurity, planned cyber-attacks, unplanned cyber-attacks, risk analysis techniques, risk likelihood

1. Introduction

The global transition due to the Covid-19 pandemic forced all institutions, including the small business sectors, to rely on Information and Communication Technology (ICT) for performing their daily activities and delivering services (Ncubukezi, Mwansa & Rocaries, 2020b). Every business sector is faced with increased risks which can be related to economic, technological, network, device, and human factors. The operators behind the increased use of technology are people who form part of the Information System (IS). People present employees and criminals who strive to gain access and work on systems (Japertas & Baksys, 2018). However, their interest in the high usage of ICT resources exposes businesses to ultimate planned or unplanned data breaches. People, as human factors, can generate cyber-risks which are two-fold: risks generated by criminals and risks caused by the employees as insiders (Ncubukezi, 2022a). The criminals exert their energies to exploit vulnerable and weak systems by gaining unauthorized access. Various reasons trigger this action. The employees as insiders also become agents of unplanned attacks, leaving room for intruding unauthorized access. Even though unplanned attacks can be caused by both insiders and criminals, the paper focuses on attacks caused by insiders.

The risks committed by the criminals are thoroughly planned, and internal risks triggered by employees or staff are unplanned. However, regardless of the source of the risks, businesses continue to be vulnerable to all kinds of cyber threats and cyber-attacks. As a result, businesses that do not have a strong cybersecurity plan in place and perform proper risk analysis for effective decision-making (Wang & Neil, 2021) become victims of the common attacks.

1.1 Research aim and objectives

This study demonstrates the risk likelihood of planned and unplanned cyber-attacks caused by human factors using the AgenaRisk package and the risk assessment techniques. To achieve this, the study:

- Identifies the causes of the planned and unplanned cyber-attacks;
- Determines the risk likelihood of the planned and unplanned attacks using quantitative risk analysis such as sensitivity analysis, scenario analysis, and decision tree analysis; and
- Demonstrates observations of the planned and unplanned cyber-attacks resulting in the risk likelihood.

The paper covers related literature about planned and unplanned cyber-attacks and threats, BN tools, related studies followed by the method of inquiry used, results, discussion, significance, and conclusion.

2. Background to the study

Regardless of the size of the business, all organisations suffer and are exposed to a range of planned and unplanned cyber-attacks (Khosravi & Ladani, 2020). The sudden Covid-19 pandemic caught institutions off guard, which resulted in a huge transition of operating from homes. That adjustment introduced remote access to the

system. The change from office to home forced employees to be independent and operate at the same homes, regardless of their level of computer skills (Khan, Brohi, Zaman, 2020). To some institutions, the change became easy due to their preparedness and readiness to use technology. The Covid-19 pandemic also became an opportunity for cybercriminals to hit all institutions and has had a negative impact on the global economy (Wijayanto & Prabowo, 2020). Even though the computer and network as the ICT were present, the issue of planned and unplanned cyber-attacks remain unclear. Planned attacks are initiated by intentional criminals that are after the lucrative benefit of any size and kind of business institution (Hakak, et al., 2020). On the contrary, unplanned attacks are unsafe and ignorant actions performed on the system by legitimate employees. All the negative and unsafe actions performed may expose the system to various negative risks.

2.1 Risks among small businesses

Risks always exist in every organisation or institution, and they are inevitable. The global Covid-19 pandemic also increased the risk likelihood in all businesses and institutions. Likewise, during the lockdown, people panicked and developed anxiety and stress. Some had to adjust from working in their comfortable office spaces to working at home. The current study focuses on the small and medium businesses described below. **Figure 1** illustrates human interaction in a networked business system with promoters of unsafe and unsecured connections. People who are legitimate employees and criminals use any type of end device to connect to the business system. With a secured connection, planned attacks are not likely to happen. The promoters of unsecured connections result in an ultimate data breach in the form of modification of information, deletion of information, and information loss. The planned or unplanned attacks that happen on the business system compromise the state of confidentiality, integrity, and availability of information (Ncubukezi, 2022b).

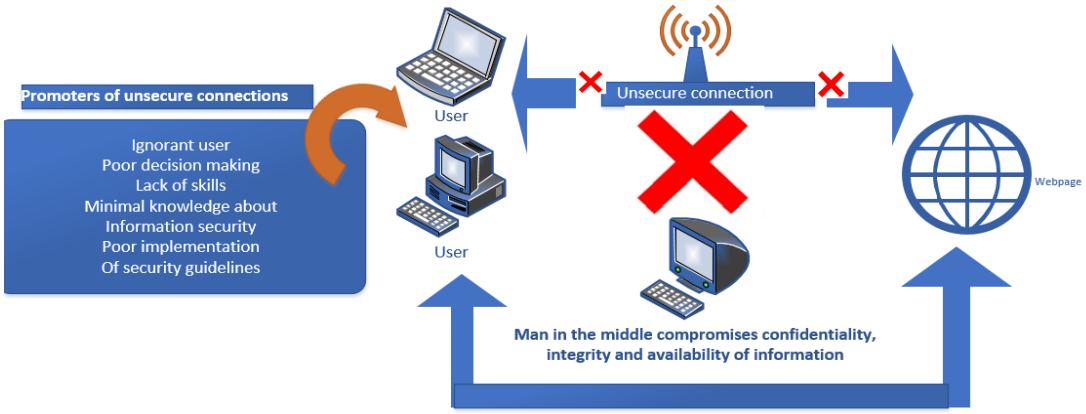


Figure 1: Promoters of the attacks in a business system (Source: Ncubukezi and Mwansa, 2021)

The attacks affect different resources of the businesses. Therefore, with the increased use of ICT among different sectors, it becomes essential to adopt and use artificial intelligence (AI), which imitates human actions and intelligence using applications and algorithms built in a computing environment (Ncubukezi, Mwansa, & Rocaries, 2020a; Fenton & Neil, 2014). These capabilities include planning, learning from experiences, and creating and adjusting input variables. This study adopted the BN, which is discussed below.

2.2 Why Bayesian Network

The study adopted the BN as the graphical model that illustrates the likelihood of the risks relating to planned and unplanned attacks. The probabilistic model denotes data about uncertain nodes holding the random variable and demonstrating conditional probability for any corresponding variable (Yang, 2019; Chen & Wang, 2017). The BN predicts the probability of risks that results from data breaches. The BN is demonstrated through the use of the AgenaRisk package, which created the interfaces (AgenaRisk, 2021). The package demonstrates, analyses, and predicts planned and unplanned risk likelihood caused by human factors. BN through the package was used to predict the risk likelihood based on the activities performed by legal or illegal users. The current scenario shows the relative probabilities of the risks, illustrating relationships between dependent and independent variables and their influence on the level of the risk likelihood (Khodakarami, Fenton & Neil, 2007).

The BN technique determines the risk likelihood based on the prior indicators, protection levels, risk likelihood, and the consequences of the risks (Fenton et al., 2007). Figure 2 shows a package that has been used to illustrate the risk likelihood of human factors such as planned and unplanned attacks. The application of the package has been demonstrated in section 4 of the study. The package interface was used to generate different scenario cases of planned and unplanned attacks.

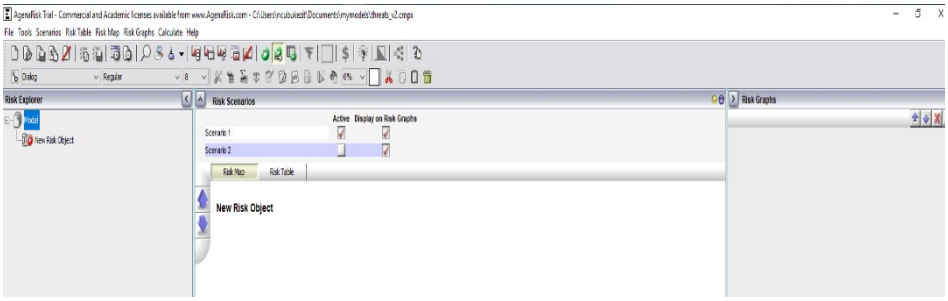


Figure 2: AgenaRisk interface

3. Method

The study collected data from the business owners and IT managers of small businesses from diverse sectors in the Eastern Cape from their natural settings, which shares the participant's activities and their actions toward their system. A total of twenty-five (25) respondents participated in the study and were selected using judgmental sampling. All the sectors that participated use the online business system to perform their daily operations, keep businesses running, and as well as increase revenues.

Data were collected using online surveys with open-ended questions. Participants were sent emails with the survey link. Collected data were categorized, analyzed, and interpreted using content analysis. The collected data was used to form the knowledge base for the development and illustration of the Bayesian Network model using the AgenaRisk technique. The agenaRisk technique is the graphical package that illustrates and determines the risk likelihood. In this study, the package determines the risk likelihood of the criminal's and employees' activities, which exposes the business to risks. Data was further interpreted using the BN through the AgenaRisk package to determine the risk likelihood of planned and unplanned cyber-attacks. The BN tool is used to support the model analysis using different analytical techniques such as sensitivity analysis, Tornado graphs.

In addition, the study used quantitative risk analysis techniques including scenario and decision tree analysis for further demonstration of the planned and unplanned attacks. Each analysis technique illustrates the uncertainty of the attacks.

The university's ethical processes, which approve research studies were complied with.

4. Results

4.1 Collected results

Data were collected from the relevant respondents who use the business systems to run their daily activities in cyberspace. Employees of the small businesses that took part in this study have a range of one to 150 per business sector and also generate a turnover of up to R1 million per year. Table 1 shows the respondent's background.

Table 1: Respondents background

Participant position	Number	Gender	Sectors	Age range
Business manager	7	Male	ICT, home improvement, retail, real estate, transport,	35-50
Business manager	5	Female		37-45
IT manager	7	Female		33-43
IT manager	6	Male		31-46

As shown on Figure 3 the selected businesses have been in existence for less than a decade. Most businesses have been operating for a year to two years, while 25% haves been operating between three to five years and 20% have been operating between six to nine years.

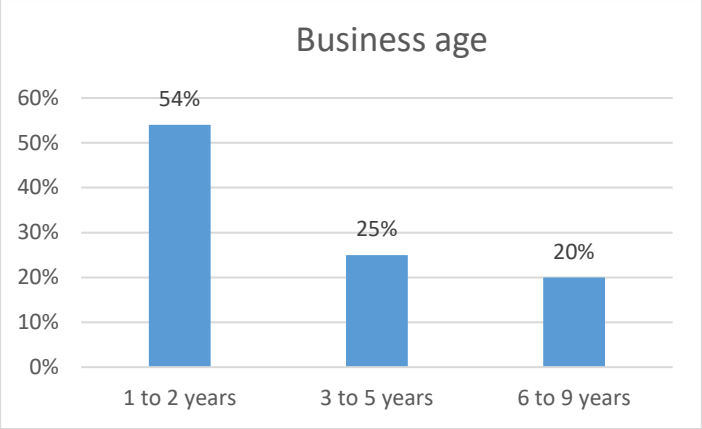


Figure 3: Business age

Figure 4 shows the causes of the planned and unplanned cyber-attacks experienced by the small business sector, due to different motivating human factors. These attacks are rated on a Likert scale from rare (1), unlikely to happen (2), moderate to happen (3), likely to happen (4), and certainly to happen (5). The respondents shared their insights about the main causes of unplanned and planned cyber-attacks.

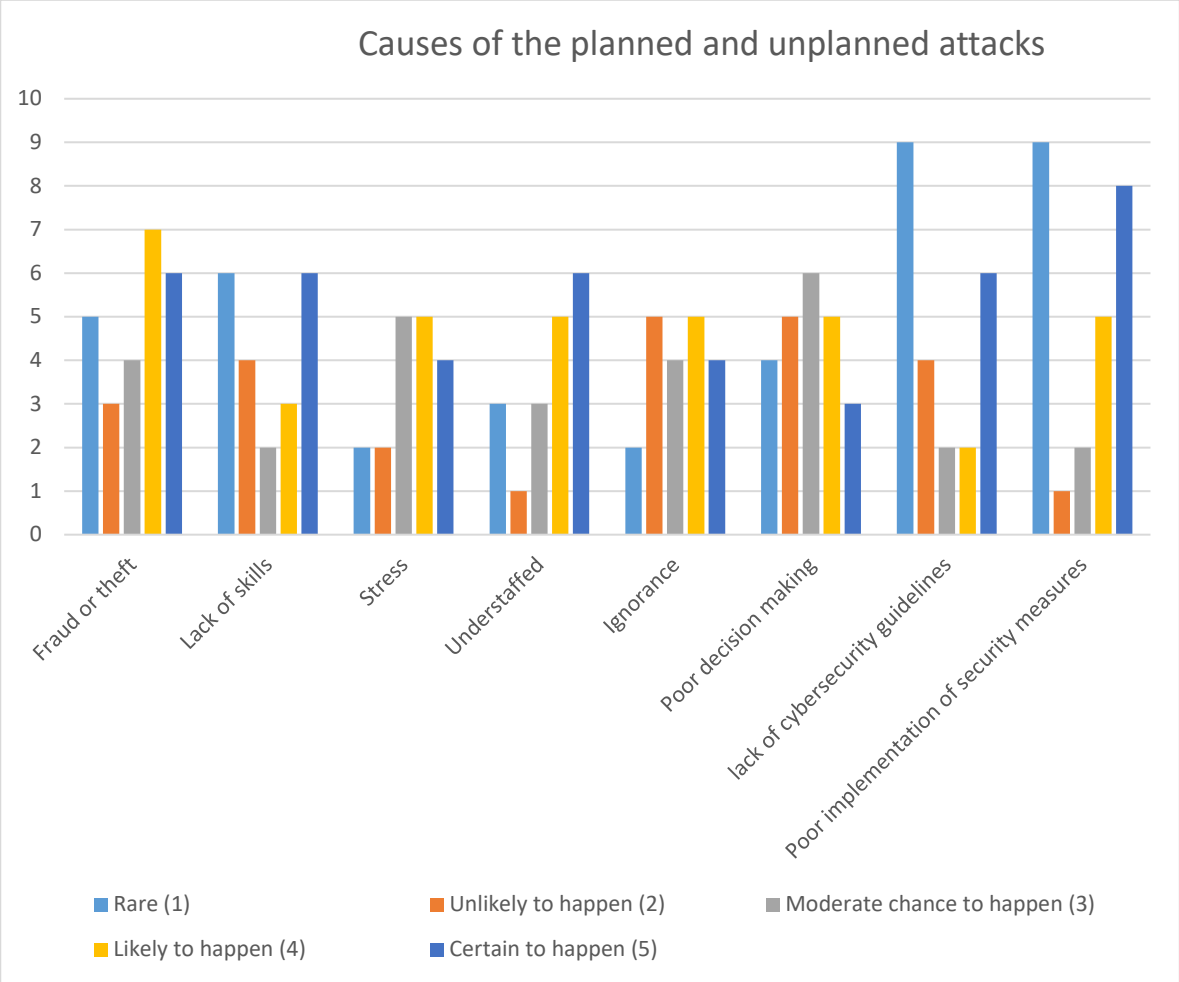


Figure 4: Planned and unplanned cyber-attacks

Results also showed different interpretations of the protection level of the businesses. Some businesses indicated that they only rely on the anti-virus as the protection measure while 100% believe in both passwords and anti-viruses.

4.2 Simulated scenario analysis

Simulated scenarios are an investigation of a particular phenomenon within its real-life context using different sources of evidence (Robson, 1993). Scenario analysis helps to separate and identify the main trends (certainties or uncertainties) by looking at the trends that may not or may be significant or may not change. The action avoids frustration and improves efficiency while saving time. The main focus is on the events involving attackers and cyber criminals, and employees demonstrating their influence on the system.

A simulated scenario of human factors was developed with two system users: a normal user and an experienced IT user. The normal user presents as an employee who is not well trained and informed to use the system, while an IT user is skilled and well-orientated towards cybersecurity best practices. In this scenario, both users can experience planned or unplanned attacks based on their actions, attitudes, and behaviours on the system. A planned attack presents a criminal looking for opportunities to gain unauthorised access to unsecured systems.

So the package predicts the risk likelihood based on different scenarios influenced by user activities and the state of security (system's protection level). The risk probability influences the risk impact. This scenario is presented owing to the high demand for Internet use and working off-site during the global pandemic, where data became the leading lucrative benefit in all organisations. Criminals always have the intention of gaining unauthorised access, while employees can commit unplanned actions which ultimately compromise the safety and security of the system. The following section demonstrates the prediction of the risk likelihood based on the different scenarios of the planned and unplanned attacks experienced by employees. An employee could be a normal system user or skilled Information Technology (IT) user who interacts with the system

4.2.1 Scenario analysis setting

These simulated scenarios are based on small businesses from diverse sectors that are mainly using technology to render services to clients. The work evaluates the cause of risk, the likelihood of the risk occurring, and its impact. This scenario describes, explores, and explains various elements that result in risks that ultimately compromise information security. They demonstrate the relationships between dependent and independent elements, the likelihood of risk, protective measures are taken, and the consequence of the risk. Scenario analysis pays more attention to the effect of risk in a certain situation. This analytical method uses specific information for certain scenarios to change the model's variables. Nodes used in the study are either ranked High, Medium, and Low. Some have the Boolean values of Yes or No and True or False. All these nodes have values that predict the likelihood of the risk and its impact.

4.2.1 Illustration of the risk likelihood on AgenaRisk

Figure 5 shows a scenario of an IT user interacting with the system. This employee experiences both planned and unplanned attacks. The planned attack could be caused by the level of employee ignorance, poor decision-making, or working under pressure. As illustrated, the planned attack has a Boolean of True and False where the prediction is 100% present. The same user also experiences an unplanned attack that could be coming from the criminals and has a Boolean of 50% False and 50% True. This means that there is a threat detected on the system which could compromise the state of security. So the risk likelihood is determined by the source of the attack and the protection level of the system ranked as High, Medium, and Low. The protection level has 50% dominating as high protection. The probability of risk is then predicted as 93.75% risk likelihood on a ranked node with High, Medium, and Low. The risk impact is then predicted as 51% high on a ranked node.

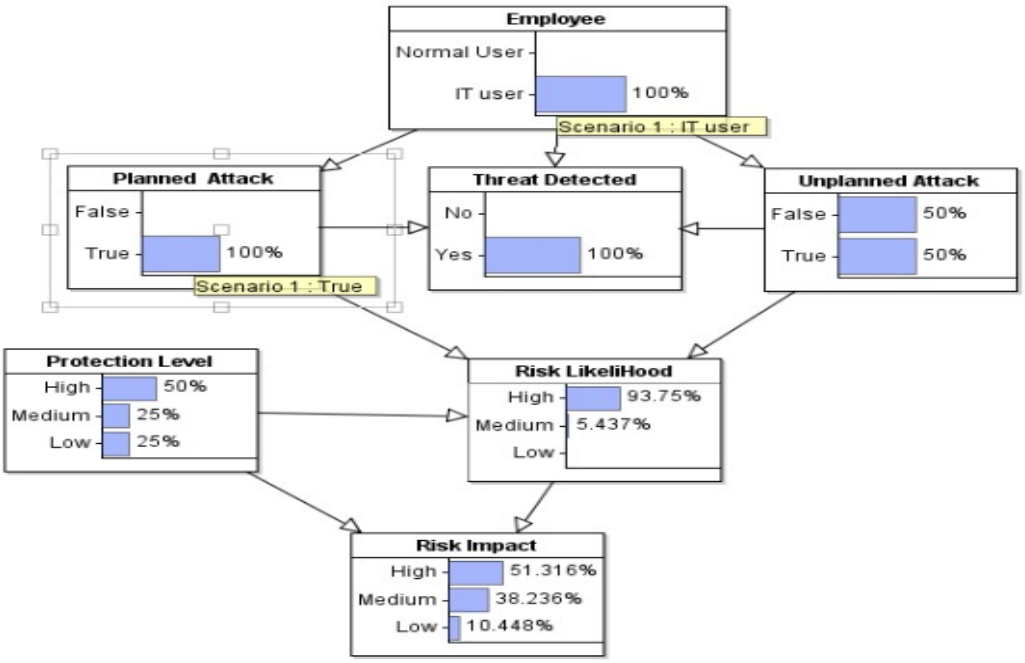


Figure 5: IT user experiences a partially unplanned and planned attack at a low protection level

Figure 6 shows an IT user with a detected threat level of 100% who experienced both 100% planned and 100% unplanned attacks. The level of attack and the partial level of protection of 50% in this scenario results in a 98% of risk likelihood. The risk likelihood and the protection level directly affect the risk impact.

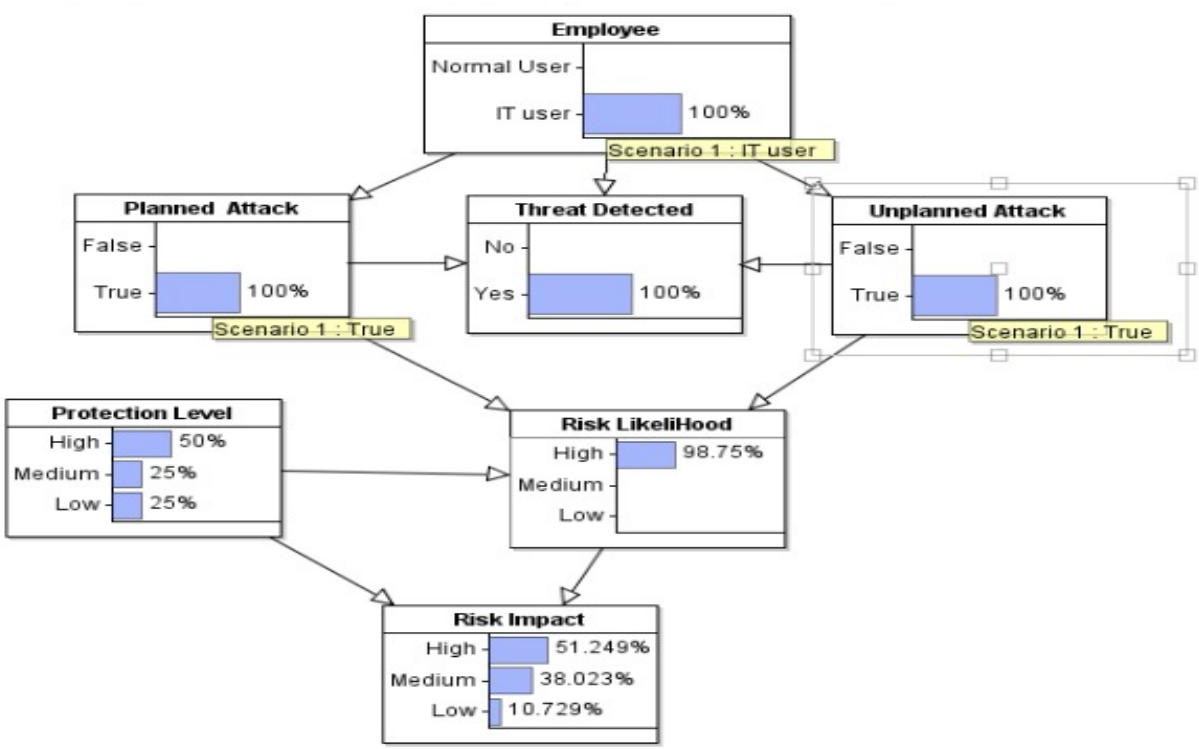


Figure 6: IT user experiences both planned and unplanned attack at the partial protection level

Figure 7 shows a normal user who experienced a 100% unplanned attack with a 100% planned attack with a 100% threat detected. These attacks happened at the 50% protection level, resulting in 98% risk likelihood and 51% risk impact.

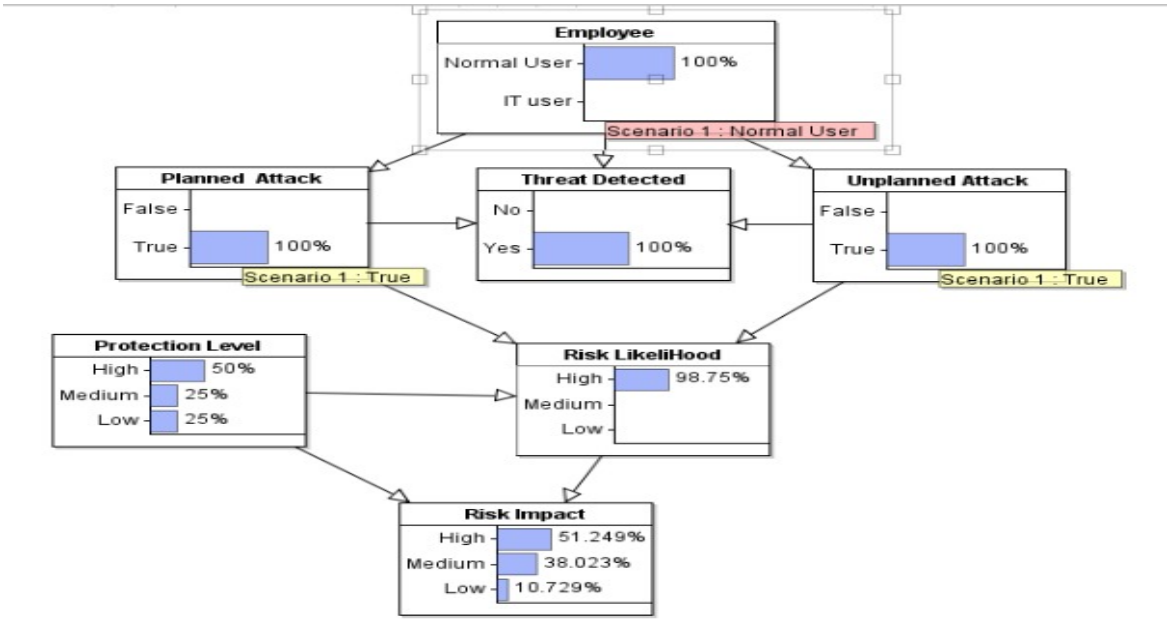


Figure 7: Normal user with 100% planned and unplanned attacks at a 50% protection level

Figure 8 shows a normal user with 68% planned and 65% unplanned attacks which means the threat is detected. The scenario shows a 100% low protection level, resulting in 96% of high-risk likelihood and 69% of high-risk impact.

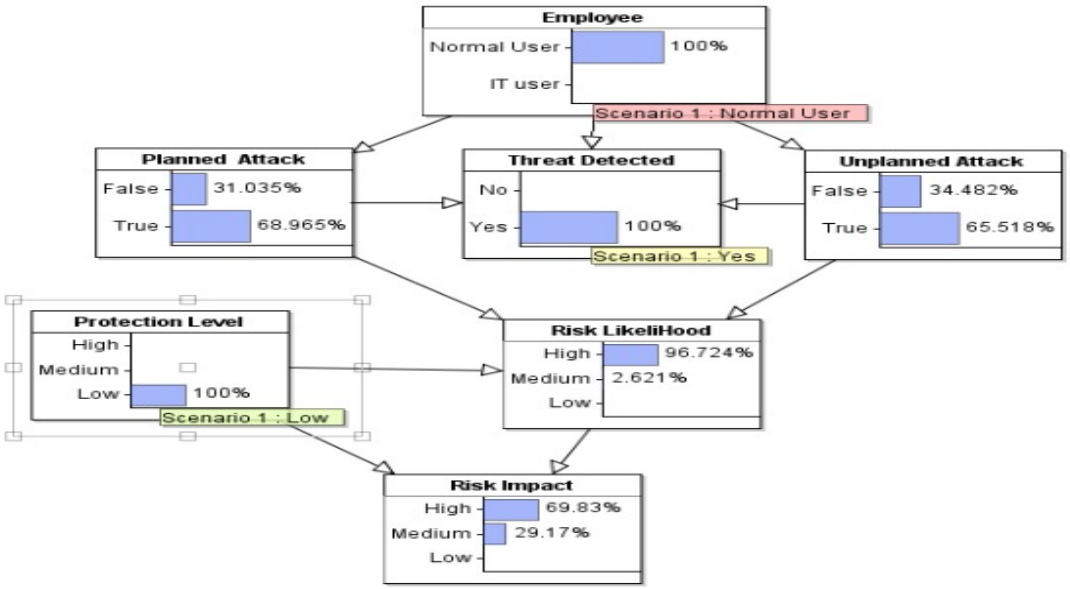


Figure 8: Normal users experience partial attacks at a low protection level

Figure 9 shows a normal user with 100% threat detected in 68% of the planned attack and 65% of the unplanned attack at 100% high protected level, resulting in 87% of the risk likelihood and 40% of the medium and high-risk impact.

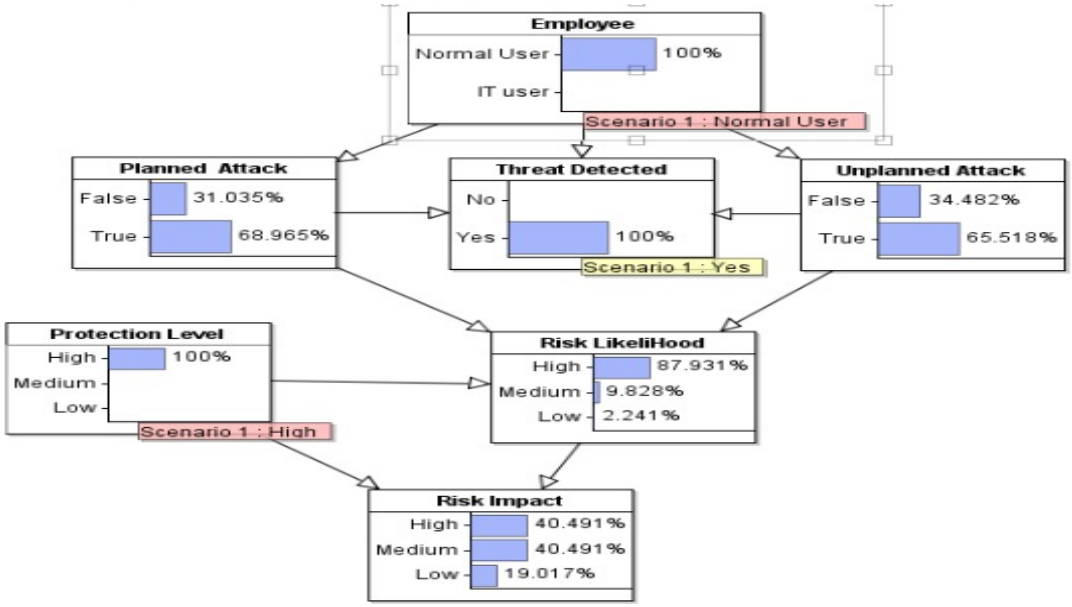


Figure 9: Combination of planned and unplanned attacks on normal user

The scenarios above illustrate both computer-literate and illiterate users who interact with the system. In all the scenarios, the user has a threat detected which could likely compromise the state of the business system safety. Even though there are all these scenarios, there is some level of security, the system remains vulnerable to both the criminals and the insiders who are legitimate employees. The absence of a dedicated cybersecurity manager could result in poor implementation of security measures (Ncubukezi, Mwansa & Rocaries, 2020b). Ignorance of internal employees caused by their level of awareness, and poor decision-making can directly contribute to the compromised business system (Ncubukezi, 2022a). The presence of both planned and unplanned attacks is caused by actions as a result of the human factors which criminals always look for an opportunity to benefit from.

4.3 Sensitivity Analysis

BN tools with the package simulated the potential cybersecurity risk probabilities and their impacts in the business space. The sensitivity analysis method examined the main causes of the risk likelihood as the results of the planned and unplanned attacks. The sensitivity analysis is performed through AgenaRisk with BN tools by selecting a single target node, and the sensor nodes are preferred based on the rank of sensitivity nodes (AgenaRisk, 2021). In this study, Boolean nodes with two values (either true or false) or ranking of (High, medium, and low) were used. Sensitivity analysis challenges the reliability, difference, and significance of the assumptions to address the 'what if' analysis. The analysis method on the scenarios checked the sensitivity of the answers against the technique and its related parameters. For the tool's effectiveness, the researcher determined the technique's sensitivity to the Conditional Probability Tables (CPT). Sensitivity analysis communicates data and outcomes, understanding the link between the input and output variables while identifying sensitive variables. It then helps to make assumptions that allow decision-making and examines the amount of risk in given scenarios.

Figure 10 shows the sensitivity analysis for the risk likelihood of a planned attack. The risk likelihood ranks from low to medium to high, while the planned attack has Boolean values (true and false). When the risk likelihood is true, it gets a high rating of the risk likelihood. When the risk likelihood is false, then the likelihood is low. The values of the sensitivity analysis are shown in the figure below.

		Planned Attack	
		False	True
Risk LikeliHood	High	0.336	0.664
	Medium	0.952	0.048
	Low	0.672	0.328

Figure 10: Risk likelihood for planned attacks (Data source: Survey, 2021)

Figure 11 shows an analysis of the unplanned attacks, which are actions performed by insiders. These attacks could be activities that are the result of poor decision-making with regard to the system of an ignorant, extremely tired, and computer-illiterate employee. Unplanned attacks hold the Boolean values (true and false) while the risk likelihood ranks from low to medium to high. The risk likelihood value becomes high when the planned attack is true and the risk likelihood becomes low when the planned attack is false. These values are shown in the figure below.

		Risk LikeliHood		
		High	Medium	Low
Unplanned Attack	False	0.563	0.427	0.01
	True	0.978	0.018	0.004

Figure 11: Risk likelihood for unplanned attacks (Data source: Survey, 2021)

4.4 Tornado graphs for planned attacks

Tornado graphs are generated based on the simulated scenario for planned and unplanned attacks. Tornado graphs are generated using the package. The Tornado chart was used to evaluate the sensitivity analysis of planned human-induced attacks. Figure 12 shows the planned attack that carries a true value with medium (0.048) and high risk (0.664) likelihood.

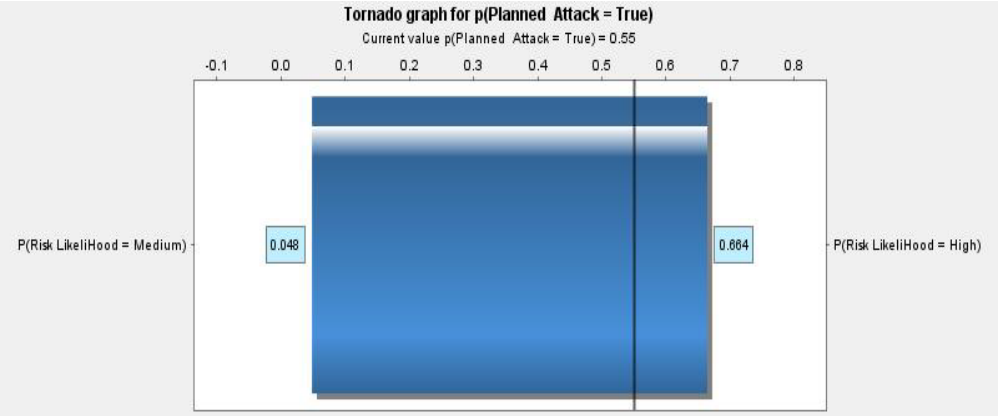


Figure 12: Tornado graph for planned attack = True

Figure 13 shows the false planned attack with high (0.336) and medium (0.952) risk likelihood.

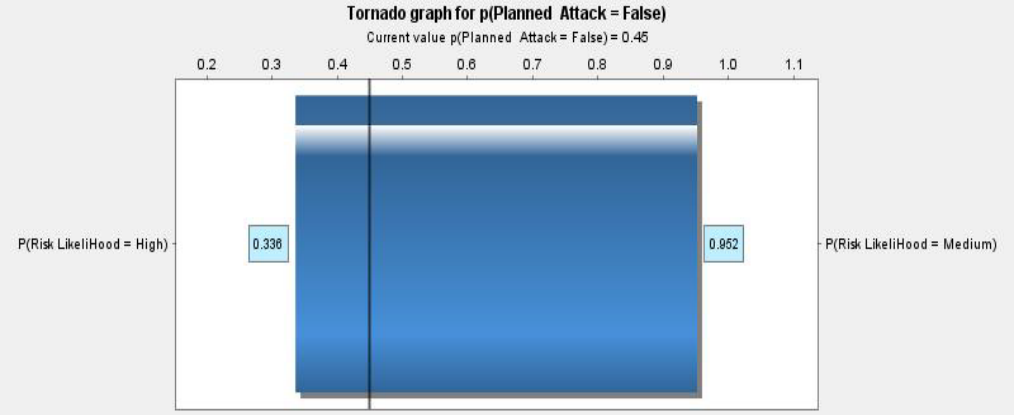


Figure 13: Tornado graph for planned attack = False

4.5 Decision Tree Analysis (DTA)

DTA as a quantitative risk analysis technique is used to assess the potential risks caused by planned and unplanned cyber-attacks. It determines the risk likelihood leading to the destruction of the business service. Every risk likelihood is determined by the appropriate and proactive security measures deployed to reduce,

avoid, separate, and mitigate risks. Figure 14 shows DTA with possible risks caused by human factors relating to planned and unplanned cyber-attacks. In the figure, people operate the system where their performed activities get processed on the business system. So these actions can be part of the planned attacks coming from intentional criminals and also from legitimate employees who sometimes ignorantly operate the business system and leave it compromised. For example, an unexpected attack could be an ignorant employee downloading a file or software from unverified sources, while the planned attack could be a criminal who has gained unauthorised access to the business system. Usually, the planned attacks are network-dependent, where a criminal searches for an unsafe network port or interface.

The planned attacks are usually in the form of phishing, malware, denial of service, and device failure attacks, while unplanned attacks relate to the lack of skills, ignorance, poor decision-making, poor implementation, unlimited access to sensitive resources, and use of the affected external hard drives. These attacks pose risks related to the lack of confidentiality, financial loss, other losses, manipulation and deletion of sensitive data, unauthorised access to private information, damage or unavailability of a resource, and delayed services.

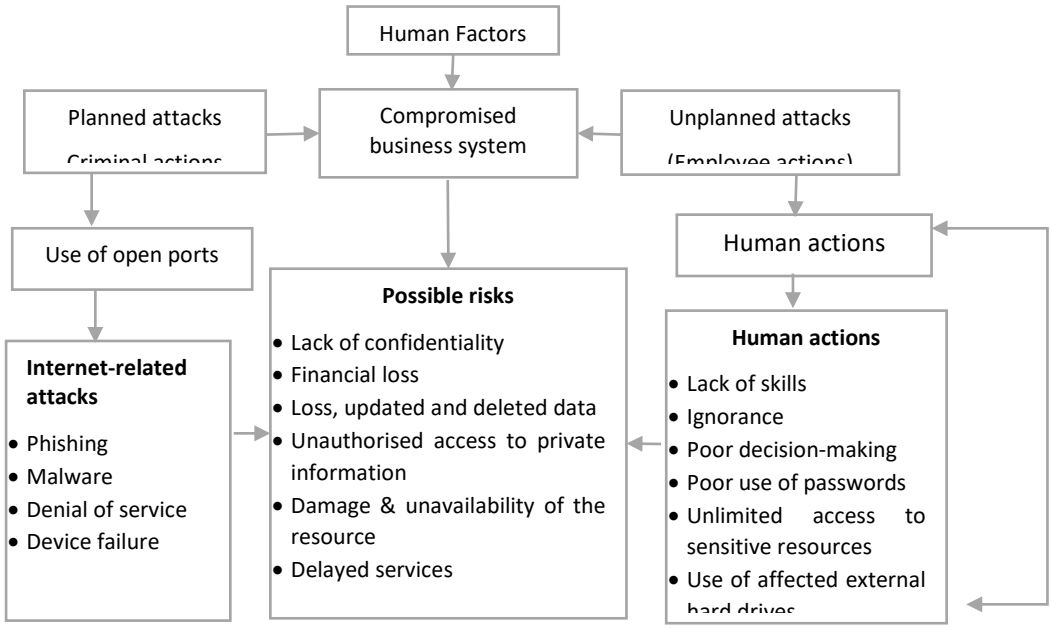


Figure 14: Risks associated with planned and unplanned attacks (Survey, 2022)

4.6 Discussions

In South Africa, there is not much deployment of artificial intelligence that replaces human actions, which makes businesses rely more on people. Even though people form a significant part of businesses, they still become the main source of various planned and unplanned cyber-attacks. Hadlington (2017) states that human errors are one of the attacks that are underestimated and ignored. The case scenario demonstrated that human factors could be the source of attacks compromising the system in one way or another. Some harmful human actions are caused by insiders, while intentional errors compromise the systems. Sometimes, human mistakes expose businesses to risks that significantly affect information and computer security (Anwar et al., 2017).

Criminals take advantage of gaining unauthorised access to the system owing to insider action. Criminals usually take time to plan for their activities, while insiders react based on their attitudes, behaviour, and opportunities. The ignorant, pressured actions reveal valuable, sensitive information and other business resources to opportunistic cybercriminals. Criminals hijack unsecured sessions intending to violate privacy and security. Minimal use of safety measures increases the chances of data breaches (Ncubukezi, 2022a). Sometimes, human errors are caused by inadequate staff members, fatigue owing to work overload, and operating under pressure (Turk, 2013).

Similarly, insiders create a loophole in the system when they have bad intentions. Understaffing in businesses creates human errors because some employees may be highly ignorant of their actions in the system. Some employees do not have adequate skills to help them make informed decisions. In this scenario, the risk probability is influenced by the presence and adherence to the available security policies, rules, procedures, and guidelines. The amount of risk likelihood influences the risk impact in the system. The higher the risk probability, the greater the impact. As stated in the previous section, cybersecurity risk results in reduced profit, poor

performance, loss of clientele, and a bad reputation, affecting overall business continuity (Ncubukezi, 2022c). Ncubukezi (2022a) states that employees can be the weakest link of the chain and a channel for criminal activities, as the criminals use every opportunity they have. A malware attack simulated scenario is presented below. Employee mistakes put businesses at risk, while the SME sector pays little attention (Ergen, Ünal & Saygili, 2021).

Often, criminals gain unauthorised access owing to employee work overload and other related factors. The challenge is the diverse range of human errors, which open doors to unauthorised access to people who steal sensitive information and other valuable assets. This action results in a significant data and security breach (Richardson et al., 2020). For example, criminals take advantage of the unsure network sessions by violating safe surfing and privacy (Wallace et al., 2020). Some threats are harmful malware such as viruses, phishing, Trojan, adware, spyware, and worms (Karaci, Akyüz, & Bilgici, 2017). For example, the software automatically gets installed without the user's knowledge when a victim ignorantly clicks on harmful links or uses automatic pop-up messages.

Table 2 shows the summary of the collected risks caused by human factors. The respondents revealed different interpretations of the protection level of their systems. Even though most businesses have indicated some awareness and understanding of the protection level of the system, none of the businesses mentioned the use of multi-factor authentication. The promoters of human errors are ignorant users who cause poor decision-making and show a lack of skills, and minimal knowledge about information and computer security, leading to poor implementation and not adhering to security guidelines.

Table 2: Summary of collected cyber-risks table (Data source: Survey, 2022)

Human factors = (Planned and unplanned cyber-attacks)					
Causes of attacks	**Affected asset**	**Security principle**	**Risk cause and source**	**Risk Impact**	**Risk consequence**
• Poor decision-making • Lack of management involvement • Work overload & stress • Ineffective access & resource management. • Outsourced IT services	• Employees • Files • Websites • Third parties • Personal information	• Confidentiality • Integrity	• Planned and accidental actions • Ignorance - poor decision-making • Lack of skills • No management involvement • No cybersecurity personnel • Poor compliance with guidelines and procedures • Low-security awareness and understanding • Poor password creation and management • Unauthorised use of access rights • Incomplete configuration management • Lack of skills	• Loss of information • Lack of client trust • Financial loss • Declined production and business growth	• Lack of a cybersecurity policy • Take your own device to work • Physical security • Identity theft, fraud. • Deletion, theft, and corruption of data.
• Poor configuration management • Malicious and accidental activities • Careless handling of data • Unsecured network ports	**Measure:** Adherence to business and cybersecurity policies and guidelines	**Measure:** Protect the • Network • Password • Device • Email • System		• Mistakes leading to a significant data breach • Ignorance can lead to legal fines, loss of client trust • Bad Reputation	• **Ultimate** data breaches

4.6.1 Significance of the study

This study shares insights about the literature relating to the state of cybersecurity and the risk likelihood of planned and unplanned cyber-attacks in small businesses. The work used the BN tools in the AgenaRisk package to demonstrate the predicted risk likelihood of planned and unplanned cyber-attacks, which is a human factor. In addition, the research used quantitative risk analysis techniques. It provides risk managers with benchmarks that can help adjust security resources in order to position ideal protection levels. This study provides risk managers with benchmarks to help adjust security resources to position ideal protection levels. The following section addresses the future developments and the concluding remarks.

5. Concluding remarks and Future developments

This study determined the risk likelihood relating to human factors which are planned and unplanned activities in the business system. The study used BN on the package to observe the planned and unplanned cyber-attacks on the business system. The observations of the risk likelihood on the package were done based on the

sensitivity analysis and Tornado graphs. In addition, decision tree analysis and scenario-based analysis were used as the techniques to analyze and determine the risk likelihood of cyber-attacks.

In the future, the observations from this study can be extended to big organisations and also be performed per sector. The scope of the work can also include other different provinces in South Africa.

Acknowledgments

I would like to thank National Research Foundation (NRF) Black Academics Advancement Programme (BAAP) Grant for financially supporting this work.

References

AgenaRisk. 2021. AgenaRisk. Available: http://www.agenarisk.com. [Accessed: 13 November 2020]

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, pp.437-443.

Chen, T.T. and Wang, C.H., 2017. Fall risk assessment of bridge construction using Bayesian network transferring from fault tree analysis. *Journal of Civil Engineering and Management*, *23*(2), pp.273-282.

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), pp.210-210.

Fenton, N. and Neil, M., 2014. Decision support software for probabilistic risk assessment using Bayesian networks. *IEEE Software*.

Fenton, N., Neil, M., Marsh, W., Hearty, P., Marquez, D., Krause, P. and Mishra, R., 2007. Predicting software defects in varying development lifecycles using Bayesian nets. *Information and Software Technology*, 49(1), pp. 32-43.

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.

Hakak, S., Khan, W.Z., Imran, M., Choo, K.K.R. and Shoaib, M., 2020. Have you been a victim of COVID-19-related cyber incidents? The survey, taxonomy, and mitigation strategies. *Ieee Access*, *8*, 124134-124144.

Japertas, S. and Baksys, T., 2018. Method of early staged cyber-attacks, detection in IT and telecommunication networks. *Elektronika ir Elektrotechnika*, *24*(3), pp. 68-77.

Karaci, A., Akyüz, H.İ. & Bilgici, G. 2017. Üniversite öğrencilerinin siber güvenlik davranışlarının incelenmesi. *Kastamonu Eğitim Dergisi*, 25(6), pp. 2079–2094.

Khan, N.A., Brohi, S.N. & Zaman. N. 2020. Ten deadly cyber security threats amid COVID-19 pandemic. *TechRxiv Powered by IEEE*: pp. 394–399.

Khodakarami, V., Fenton, N., & Neil, M. (2007). Project Scheduling: Improved Approach to Incorporate Uncertainty Using Bayesian Networks. *Project Management Journal*, 38(2), pp. 39–49.

Khosravi, M. and Ladani, B.T., 2020. Alerts correlation and causal analysis for APT based cyberattack detection. *IEEE Access*, *8*, pp.162642-162656.

Ncubukezi, T. and Mwansa, L., 2021. Best Practices Used by Businesses to Maintain Good Cyber Hygiene During Covid-19 Pandemic. *Journal of Internet Technology and Secured Transactions*, *9*(1), pp. 714-721.

Ncubukezi, T., 2022a. Human errors: A cybersecurity concern and the weakest link to small businesses. In *Proceedings of the 17th International Conference on Information Warfare and Security* (p. 395).

Ncubukezi, T., 2022b. Impact of information security threats on small businesses during the Covid-19 pandemic. In *European Conference on Cyber Warfare and Security*, 21(1), pp. 401-410.

Ncubukezi, T., 2022c. *Design development and evaluation of the cybersecurity risk tool: a case of small and medium-sized enterprises in South Africa. (*Doctoral Thesis, Cape Peninsula University of Technology, unpublished).

Ncubukezi, T., Mwansa, and Rocaries, F. 2020a. "A Proposed: Integration of the Monte Carlo model and the Bayes network to Propose Cyber Security Risk Assessment Tool for Small and Medium Enterprises in South Africa." *International Journal of Computer Science and Information Security*, 3(18), pp. 152-155.

Ncubukezi, T., Mwansa, and Rocaries, F. 2020b. "A review of the current cyber hygiene in small and medium-sized businesses." *International Conference for Internet Technology and Secured Transactions*, 15, pp. 283–288.

Robson, C. 1993, *Real world research: A resource for social scientists and practitioner-researchers*. Oxford: Blackwell Publishers.

Turk, R.W. 2013. *Preparing a Cyber Security Workforce for the 21st Century*. Army War College Carlisle Barracks, PA.

Wang, J. and Neil, M., 2021. A Bayesian-network-based cybersecurity adversarial risk analysis framework with numerical examples. *arXiv preprint arXiv:2106.00471*.

Wijayanto, H. and Prabowo, I.A., 2020. Cybersecurity Vulnerability Behavior Scale in College During the Covid-19 Pandemic. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, *9*(3), pp. 395-399.

Yang, X.S. 2019. *Introduction to algorithms for data mining and machine learning*. London: Academic Press.

Ret-gadgets in RISC-V-based Binaries Resulting in Traps for Hijackers

Toyosi Oyinloye[1], Lee Speakman[2] and Thaddeus Eze[1]
[1]University of Chester, Chester, UK
[2]University of Salford, Manchester, UK
t.oyinloye@chester.ac.uk
l.speakman@salford.ac.uk
t.eze@chester.ac.uk

Abstract: The presence of instructions within executable programs is what makes the binaries executable. However, attackers leverage on the same to achieve some form of Control Flow Hijacking (CFH). Such code re-use attacks have also been found to lead to Denial of Service (DoS). An example of code re-use attack is Return Oriented Programming (ROP) which is caused by passing input crafted as chained sequences of instructions that are already existing as subroutines in the target program. The instructions are called gadgets and they would normally end with ret. The ret instructions enable the flow of hijacked execution from one set of instruction to another within the attacker's control. There could however be exceptions depending on the structure of the chained gadgets where the chained gadget fails to run its course due to inability of specific gadgets to replace the value in the return address (ra) register. The dangers of chained gadgets are not a new idea but the possibility for an attacker's gadget chain to fall into a trap during a ROP attack is not commonly addressed. In addition to this, recent studies have revealed that understanding the behaviours of gadgets would be useful for building information base in training machine learning (ML) models to combat ROP. This study explains the behaviour of certain ROP gadgets showing the possibility of occurrence of a loop in execution during exploitation. A sample program which accesses gadgets from the GNU C library (glibc) is used to demonstrate the findings. Gadgets identified with this possibility are poor for chaining as they do not contain instructions to load or move new values to the ra register and would produce unreliable exploits. This would result in a trap for the chained gadgets instead of arbitrary code execution, and DoS on the path of the user. This implies that the impact that a ROP chain could have on a targeted process does not only rely on the underlying system architecture but also on relies on the structure of the chained gadget. In this paper, the RISC-V architecture is the focus, new gadget finders (scripts are available) are presented, and sample of chained gadgets are analysed on a RISC-V - based binary.

Keywords: RISC-V; Return Oriented Programming; Denial of Service; Ret-gadgets; Code-reuse; Control Flow Hijack

1. Introduction

Ret-gadgets are the weapons of warfare in ROP (Shacham, 2007) attacks and as long as vulnerable programs remain, it is important to seek ways of providing protection for vulnerable programs. In the effort to protect vulnerable programs, it would be valuable to further examine the weapons of attack based on their underlying architectures and behaviour of the attack vectors. This would empower stakeholders to better understand what they are fighting against. Attackers of software have developed various techniques for CFH. While buffer overflow remains a fundamental route to complex attacks, further exploitations are made feasible due to executable instructions in the programs. Existing code content in executables offer functionality to the program but on the other hand, could become dangerous tools in the hands of hijackers. Existing code ending with ret referred to as gadgets are used by attackers to circumvent basic protection techniques like DEP/NX (Microsoft, 2018) and ASLR (Pax Team, 2003). The gadgets reside as subroutines in the text or shared library section of the original program. As the gadgets lie within, they can be accessed throughout the runtime of the program and are ready tools that attackers use for code-reuse attacks.

Code reuse attacks involve the use of existing code within executables to perform tasks that are different from the original intention of the programmer. Code reuse attacks could occur as CFH via buffer overflow which escalates into ret-to-libc and then ROP. The ret-to-libc could occur when attackers pass addresses of existing functions as input into a target while ROP could begin and end successfully with a chain of gadgets ending with ret, jump or a call as input via buffer overflow. A malicious input could be used in a buffer overflow followed by redirection of control flow to an injected malware or to an unintended destination within the target program. If the attacker passes an input that only contains addresses and values, they can achieve a ROP attack. The first address redirects execution to a different instruction and once that instruction is executed, it returns to fetch the next address to be executed until a full-blown ROP is achieved. A ROP chain would not always lead to a complete ROP attack but would still inhibit the purpose of the running process. This is because the completion of a ROP attack depends on the level of complexity of chaining the gadgets, as well as the behaviour of the gadgets that are included in the ROP chain. Availability of gadgets within a binary depends on the way the

system architecture interprets the executable binary. The forms of the gadgets would then determine the possible variations in sequences of chained gadgets and ultimately the outcome of the ROP attack.

The possibility of complete ROP on different architectures has been demonstrated in various studies but the gadgets that are used in achieving the ROP are not widely studied particularly for the RISC-V architecture. For example, studies (Shacham, 2007; Checkoway et. al, 2010, Bletsch et al., 2011, Carlini & Wagner, 2014) on implementation of ROP attacks and possible means of mitigations for ROP (Abadi et al., 2005; Nui, 2014)have been presented with regards to programs built and run on the X86 and ARM CPUs. This is justifiable as X86 and ARM have been the commonly implemented systems for computer devices and smart devices respectively. In recent times, RISC-V, a new CPU architecture, has been developed and is fast gaining popularity particularly among producers of electronic voting machines, smart devices, personalised health management systems, etc. (IIT Madras, 2020), and recently laptop computers (RISC-V Community News, 2022). While the possibility of ROP across different architectures has been established, the complexity of ROP chaining, the variety of possible outcomes, and behaviour of ROP gadgets particularly on the RISC-V-based binaries are not yet widely explored. For this reason, this study focuses on the RISC-V architecture. Understanding the behaviour of various gadgets would be valuable for predicting the feasibility of a ROP attack on any architecture.

Gu and Shacham (2020), and Jaloyan et al. (2020) revealed that the underlying architecture of the executable program could impact the feasibility of ROP. The CPU interpretation of instructions could generate new instructions that could be useful as gadgets but not detectable by existing protections. In addition to this, the existing protections have not reasonably considered vulnerability of RISC-V-based binaries. Apart from the knowledge of underlying dangers of ROP on RISC-V, recent study by Koranek et al. (2022) also presents models that were developed with focus on ROP gadgets as valuable resource for ML towards detection of possible ROP. Understanding of the behaviour of gadgets would be useful in this regard.

A sample C program compiled in RISC-V was examined and two new gadget finders were built in the process of this study. The program is then exploited as the target for ROP attack with focus on gadget *getpid* which exists in glibc. The C program can access glibc during execution which indicates that the event presented here can be replicated in any other Linux on RISC-V-based binary. In addition to this, we highlight how the characteristics of ret instructions may vary as they tend to behave in line with the operations that they were originally built to perform. Secondly, we highlight that the addresses that are stored as return address require to be overwritten with the next destination in a ROP execution on the RISC-V-based binaries. The form of ret instruction that is encountered in the ROP chain would determine the next step that the ROP execution would take. This could be a success and progressive ROP attack or a failed attack which results in a trap for the hijacker, and a DoS for the user. With reference to these findings, we demonstrate entrapped ROP on the target. The outcome of this study is valuable information that could be harnessed for building ML technique towards detection of malware and CFH.

As software exploitation via ROP is an on-going challenge in software securities, existing mitigations for DoS caused from ROP is discussed. The remainder of this paper is structured as follows: Section 2 discusses related works. Section 3 presents details on the approach adopted in this study. Section 4 discusses the selected ret gadgets that were extracted from the sample program and our implementation of ROP using the gadgets. Section 5 discusses the findings. Section 6 is a conclusion and brief on future works.

2. Related works

There are previous studies which have discussed the use of ret gadgets for software exploitation in form ROP via CFH. One of the major studies around the concept of gadgets that resulted in turing complete ROP was presented by Shacham (2007) where ret-to-libc and Return Oriented Programming(ROP) were demonstrated. Shacham's study was demonstrated on the X86 architecture with the use of large combination of short sequences of instructions to achieve a ROP, while substantial tasks towards ret-to-libc were achieved using whole blocks of functions existing in libc library. Shacham's gadgets were extracted from X86 executable program through static analysis and using Galileo's algorithm. Similarly, other studies have presented variations of ROP on X86 in form of Jump Oriented Programming (Bletsch et al., 2011) and Call Oriented Programming (Carlini & Wagner, 2014)). While the gadgets used by Bletsch et al. consist of sequences of indirect jump instructions, gadgets used by Carlini and Wagner consist of instructions that end with indirect calls.

This study adopts methods from Shacham (2007), first by using static analysis via objdump tool and other Linux commands that are useful for static analysis, to fetch available gadgets but in addition, dynamic analysis via gdb was used to enable a view of the process during the implementation of the ROP exploits. We highlight the

behaviour of specific gadgets which are available in glibc to demonstrate ROP based on chained sequences of instructions on a different CPU, which is RISC-V. Similarly to Bletsch et al., (2011) and Carlini and Wagner (2014), the gadgets in this study include jump oriented gadgets and gadgets ending with indirect calls.

Study by Checkoway et al.(2010) presents implementation of ROP using gadgets without returns on X86 and ARM architectures. According to Checkoway et al., there are two properties that make it possible for return instructions to be misused as gadgets in chained malicious input. First, the possibility of transferring control of execution via indirect jumps, and secondly, their ability to update processor state such that subsequent flow of execution is not transferred back to the same location. This study agrees with Checkoway et al.'s statement on these properties as a violation of the second property would effectively cause a loop in the execution which we consider in this study as a trap for the hijacker's malicious code. Apart from entrapping the ROP, it is also observed that this would lead to a DoS for the user. Checkoway et al. presented instruction sequences that behave similarly to returns. This was achieved by inserting gadgets that contained instructions that could act as trampoline for the ret without return instructions.

In this study, a similar method to that of Checkoway et al. (2010) is applied through the use of sequences of instructions called chargers or linker gadgets. The gadget that would then carry out the needed task of full exploitation is termed as functional gadget. In the implementation done in this study, the order of the sequence of instruction was a deliberate plan of inserting the *no-return* gadgets within the gadget chain to show how the loop is established. The terms chargers, linkers and functional gadgets were adopted from the theoretical plan for ROP on RISC-V by Deac (2022). Also, while Checokway et al., made use of the same algorithm as Shacham, the gadgets in this study were extracted using newly written scripts for use within RISC-V system.

Other similar works that studied exploitation of executable binaries via ROP are Gu and Shacham (2020), and Jaloyan et al. (2020) which both revealed that RISC-V is prone to a turing complete exploit via ROP. According to Gu and Shacham (2020), this could be achieved if an attacker uses gadgets retrieved from glibc in forming ROP chains. Jaloyan et al. (2020) also identified that the underlying structure of RISC-V could be an opportunity for attackers to retrieve gadgets that are not detectable by existing gadget finders or analysis tools. For this reason, there could exist hidden gadgets in the Control flow which would not be considered in the development of existing ROP mitigation techniques. Although this study does not emphasize on this fact, it is important to note that such gadgets continue to threaten the security of software.

A recent study by Koranek et al. (2022) also focused on ROP on RISC-V to proffer solutions towards detection of ROP on RISC-V by using deep learning (DL) models to distinguish features from execution trace. Koranek et al. analysed branch patterns in ROP as valuable information towards ROP detection. They presented models that were capable of predicting the classification of execution trace thereby enabling detection of possible ROP exploits. This study relates with Koranek et al.'s goals to contribute to malware detection and detection of possible ROP exploits. According to Koranek et al., DL and ML have been found to be useful for detecting malware particularly with the complexity of malware evolution that we are currently facing. The more information that is obtained and analysed from numerous samples of programs, the more empowered the ML training, testing and validation would be. The sample shown in this paper involves code from glibc which is normally accessible to any executable running on the Linux system. The form of exploit demonstrated can be replicated on other vulnerable executable programs compiled on RISC-V system. Results presented in this study would be valuable input for building ML development/training algorithms towards detection of ROP and other CFHs.

3. Approach to the study

In this study, technical steps were taken to acquire sequences of instructions, identify viable sequences of instructions, and build gadgets chains. Further steps include exploiting the target program with the chained ROP input to record and analyse specific outcomes of passing chained gadgets. We began by compiling the sample C program that were written purposely for the study of software exploitation. The programs are vulnerable to buffer overflow and ROP. We compiled the program statically such that the libraries are included within the executable program, the optimisation was set to 0. This is to give us more access to all the available sequences of gadgets, and lots of options to choose from. The program was compiled with gcc and analysed on the platform where it was compiled.

Once the program was compiled into executable, further steps were taken towards selecting/analysing viable gadgets using two new gadget finders, deliberately chaining gadgets in specific order, and passing the chained gadgets as input into the target to mount ROP. These steps were achieved based on possibility that selected

gadgets would give an outcome that is different from the original purpose of the target. The selected gadgets were chained together tin the exploitation phase. *gdb* was used for dynamic analysis for observing execution and recording of the outcome of each exploit.

The behavioural pattern of each gadget chain was recorded highlighting specific order of chained gadgets that lead to traps for the ROP input, such that the execution continues to loop over the last bunch of instructions. It was identified that depending on the carefulness of the attacker with crafting the gadget chains, there could be three different possibilities. The input might cause the target to crash, or fall into a trap, or allow a complete execution of the instructions held in the gadgets addresses without the user knowing. All demonstrations in this paper were run on Linux RISC-V (RV64) architecture. RISC-V is not yet operational on computers. The study system is an emulated Linux Fedora RISC-V64 which was built on a QEMU emulator. The emulator resides on a Linux Fedora computer system.

3.1 Gadget finders

There were no RISC-V gadgets finders accessible during the course of the study, so we wrote two new scripts as gadget finders which made a variety of gadgets available. One of the gadget finders could find sequences of instructions ending with *ret* while the other could find sequences of instructions ending with *jal* or *jalr*. Instructions ending with ret is a general concept in execution but the *jal* and *jalr* are peculiar to RISC-V. *jal* and *jalr* both end up being interpreted as call to a function based on the return address. The gadget finders, *RETGadget* and *JALRGadgets* were written in Linux scripts specifically for RISC-V-based programs. They make use of static analysis tool *objdump* and other linux command line tools for extracting specific sequences of instructions. Extracted output get written into text files from which further analysis were done.

3.2 Extracting and Selecting gadgets

To extract gadgets from the sample program, the program was passed as argument to each of the gadget finders. Gadgets that could fit into the target frame were selected from the output files. Some of the extracted *ret* gadgets have attributes as identified by Checkoway et al. (2010) in the two conditions: (i) To transfer control flow via indirect jump, (ii) Overwrite previous return address to ensure that control flow is not transferred to the same location but some ret gadgets do not fulfil the second condition. Although they are not valuable for achieving a full ROP, they can be used to cause a DoS. These group of gadgets are the ones that result in the loop in the execution flow, whereby the rest of the gadget chain is trapped and unable to fulfil the purpose of the attacker. Aside from gadgets ending with *ret*, other variety of gadgets that were aimed for were those ending with *jal* and *jalr*. These gadgets are useful to give the attack chain more capacity and control They are useful for passing new values onto the stack to overwrite values initially held in general purpose registers.

3.3 Crafting the ROP chains

Once useful gadgets were selected, we mapped out ROP chains in various order based on the theoretical approach by Deac (2022). ROP on RISC-V is complex especially compared to ROP on x86. The stack on RISC-V is managed a bit differently and manoeuvring along the stack via a ROP chain requires lots of gadgets in the input byte stream. In addition to this, on RISC-V, *ra* (return address) register holds different values from time to time during execution. For this reason, lots of manipulations can be done to redirect the flow for as long as the *ra* register can constantly be overwritten. According to Deac (2022), the charger gadgets can be passed along with desired values in order to store new values and addresses in available registers, making it possible to achieve the ROP. The goal is to create a fake frame, particularly in form of a function epilogue. This tricks the process into assuming that the previous function has been completed and then returns to the stack to fetch the address of next instruction, in which case, execution jumps to another fake frame as crafted by the attacker. This would go on until the execution encounters an instruction that halts the execution.

The outcome that each of the gadget chains achieves depends on the intention of the attacker. For this study, some of the gadget chains were crafted such that the execution falls into a loop to show how the attacker's input is trapped, causing a DoS. Other options of chained gadgets are also demonstrated to show how the gadget chain could be executed to completion in two different categories. In all cases demonstrated in this study, two classes of gadgets classified by Deac (2022) were chained sequentially as functional gadgets, and charger gadgets. The functional gadgets will hold the instructions for the actual attack or purpose, while the chargers (linkers) gadgets will load the registers with addresses of the functional gadgets and other useful values. The linkers are the ones responsible for creating the fake frame that introduces a fake epilogue to the preceding

function, while each of the functional gadgets would be called from the new fake frame. With the functional gadget, lots of other exploits or further creation of fake frames could be achieved. For this reason, viable ROP chains for RISC-V would normally be longer than that of x86.

3.4 Passing the input

The input is passed as byte streams stored in a file. As the sequence in which gadgets are crafted in RISC-V would determine the outcome of the exploit, specific sequences that enable the outcome that are desired in the context of this study are used. This has a lot to do with the value of the *ra* register that gets overwritten from time to time as execution steps into and out of library functions or other functions that get called within a function. The branching pattern can only be detected during dynamic analysis and remain undetectable to users as no feedback is written to standard output.

4. Ret Gadgets on RISC-V and implementation of ROP

In this section we demonstrate the compilation of the sample program, extraction of gadgets, selection of gadgets, and implementation of ROP using input crafted from combination of gadget sequences formed using the selected samples. The outcome for each exploit is also discussed. This results in three categories of ROP. The first category makes use of a gadget that enable the *ra* to be overwritten as well as ending with *ret*. The second category uses a gadget ending with *jalr* instruction which could be harnessed in performing some exploits or just manipulating registers. The third category are gadgets that do not overwrite the *ra* and thus lead to a loop in the execution. The ret gadgets works more efficiently within fake epilogues but the *jal* or *jalr* can be used more flexibly. For this reason, the linker gadgets were built out of ret instructions. However, gadgets with ret that don't have an instruction to load a fake *ra*, can lead to a trap or it could end in some unexpected behaviour of the executable. The *jal/jalr* instructions, unlike the *ret* would enable the jump and link of execution to address held in the specified register. The order in which the gadgets are arranged in the chain would also affect the outcome of the exploit.

4.1 Compiling the programs

The program was compiled using gcc compiler and optimisation set to 0 to keep the code as large as can be and increase the possible available gadgets. The sample program is small but compiling it statically ensures that the libc code are also held in the resultant executable, making it larger than would normally be, and further increasing the number of available gadgets. The stack protection is bypassed in ROP, so the program was compiled with default setting including stack protections. The -g flag is also included to enable optimal dynamic analysis using gdb.

$ gcc stack_exploit_2.c –o stack_ex_s –g –O0 --static

4.2 Extracting the gadgets

The compiled executable was passed as arguments while running each of the gadget finders. The new gadget finders extracted 26 *jalr* gadgets, 89 *jal* gadgets and 1,657 *ret* gadgets. Screenshots of the extracts for *jalr* and *ret* gadgets are shown in figures 1 and 2.



Figure 1: Extracted gadget ending with *jalr*

The gadget in figure 1 begins with instruction *li a4, 0* to load an immediate value of 0 into a general-purpose register a4 which would normally be used by the process to store a function argument. The next instruction *mv a3,s0* will copy the content of register *s0* into register *a3*. Registers in the range of *s0-s11* would normally hold values that persist after function calls. *s0* is the saved frame pointer and value in there should have been used to keep track of the stack. The next three lines of instruction are *move* instructions, copying the content of each of the saved registers *s4,s3*, and *s2* into *a2, a1* and *a0* respectively. The instructions are useful for ROP if exploit values have been passed earlier on into those *s0,s4,s3*, and *s2* registers. The last line of instruction is the jump and link register *jalr s1* which would write the value in *s1* into *ra*, and then pass execution to the instruction in

there. Both *jal* and *jalr* instructions would pass new values to *ra*. *Jal* would pass a label and *jalr* would pass value stored in the stipulated register into destination register *rd* which is further interpreted as *ra*. This shows that *jal* and *jalr* gadgets are more useful as functional gadgets. As mentioned earlier, a charger gadget would be useful to control the registers before these types of gadgets can be applied as functional gadgets. The charger would preferably be a gadget ending with *ret*. An extract of gadget ending with *ret* is shown in figure 2.

```
51644:      60e2                ld      ra,24(sp)
51646:      6442                ld      s0,16(sp)
51648:      64a2                ld      s1,8(sp)
5164a:      4505                li      a0,1
5164c:      6105                addi    sp,sp,32
5164e:      8082                ret
```

Figure 2: Extracted gadgets ending with *ret*

The gadget in figure 2, ends with *ret* and begins with a *ld ra,24(sp)* load instruction. This is useful for taking control of the stack particularly the *ra* register which ultimately determines the branch pattern of the process. In other words, it can be triggered as a fake frame to trick the process into executing like a normal function. With the first load instruction, we can overwrite the value in *ra* with the address of the functional gadget that is shown in figure 1. The mapping on the stack for the value to be written to *ra* reserves a position of 24 bytes from the top of the stack for this. So that when we pass our input into the process, we must ensure that the address of the functional gadget is positioned correctly in the chain to get it stored in that location. Other values can be passed along with the new *ra*, and strategically linked to the chain to ensure that they are written into the current location on the stack which will enable the writing of new values to the registers *s0, s1, and a0*. The fourth instruction is *li* (load immediate) which loads 1 into *a0*. The line of instruction *addi sp,sp,32* deallocates the stack before the return is triggered.

4.3 Chaining the gadgets for specific goals

Several forms of gadgets were extracted. The type of gadgets available in a target program depends on the original purpose of the process itself and the CPU's way of interpreting the instructions. It is noteworthy to mention that library code is accessible to all the executables and gadgets can be fetched from there as well. Each category of gadgets that we have observed shows possible variations in the behaviour of gadgets. The *ret* gadgets, when they are used within fake frames, would have to return to address that is held in *ra*. While this is effective for establishing branching pattern, it could limit the instruction and restrict the flexibility in the manoeuvring for the attack. On the other hand, gadgets ending with *jal* or *jalr* can be used more flexibly. In demonstrating ROP here, we show how these instructions can be used with regards to their distinct advantages for changing values held on the stack and inside useful registers and then changing the behaviour of the process. Portions of the constructed input were also featured in recent work by Oyinloye et al. (2022). However, specific highlight here is on gadgets with *ret* that do not contain any instruction to load a fake *ra*, thereby leading to a trap for the input. Three examples of chained sequence of gadgets are demonstrated in the following subsections.

4.3.1 Chained sequence to create a fake frame

The ROP chain here is made of a charger and a functional gadget *exit* to show how target could be exploited with a safe exit at the end. This was passed as input that effectively hijacks execution and invokes an abrupt end. This also demonstrates the use of the linker beginning with an instruction to load a new address into *ra* register as the choice of gadget for the first element in the sequence. In figure 3, the charger creates the fake frame with values loaded or copied from one location or register to the other. Once execution is redirected, the new values are loaded into specified register *s1*, which was duplicated into *a0,* another register *s0*, and then new value for return address is written into *ra.* In this case, it points to the exit instruction.

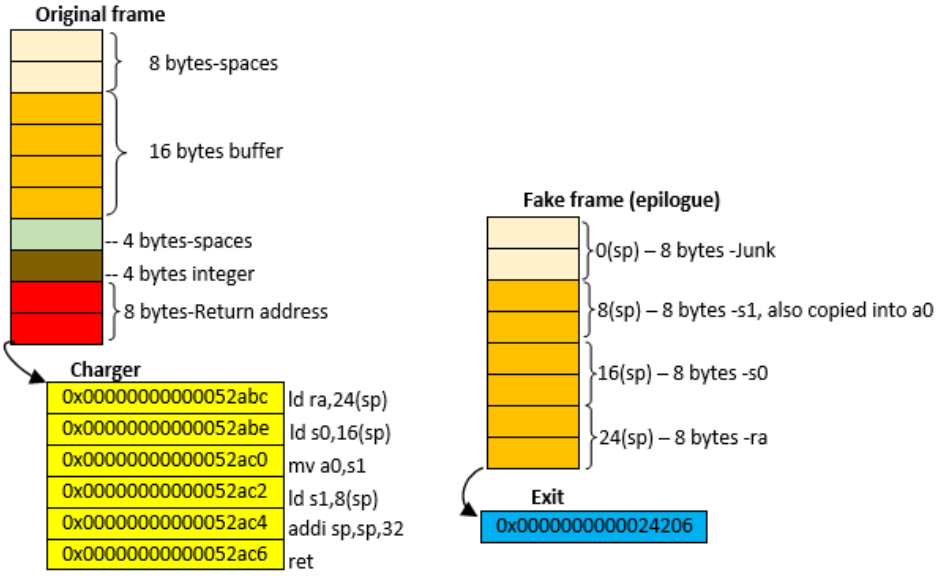


Figure 3: **Chaining gadgets to build fake frame and exit abruptly (Oyinloye et.al., 2022)**

This input in the first instance, overwrites the original *ra* with the address of the charger so that the fake frame will be running in a different address with different *sp* in the target process. Within the charger frame, a new *ra* is loaded with the address of *exit* using the ld *ra,24(sp)* instruction at the top of the charger. Other registers are loaded with junk using the charger and the input is passed into the stack of the target via the buffer. This results in an abrupt stop in execution.

4.3.2 Chained sequence to invoke a functional gadget

In this category, the ROP chain is extended by including a sequence of instructions ending with *jalr* and then an *exit*. For this example, we use a different charger. This is so that there could be more registers to manipulate by writing new values into them and then using a functional gadget to invoke the address that is written from the charger. This demonstrates how a functional gadget can be worked into the ROP to perform more exploit. The input passes values into *ra* & *s0-s2*. In the new frame, which is created by the charger gadget, *ra* will be the address of the functional gadget. *s0* and *s1* are filled with junk. The address of *exit* is then written to *s2*, to be called later from the functional gadget. The functional gadget duplicates the value in *s1* into *a0* and then *s0* into *s3* as shown in figure 4. This step is useful for the attacker to pass values into the registers that they might find useful for the exploit.

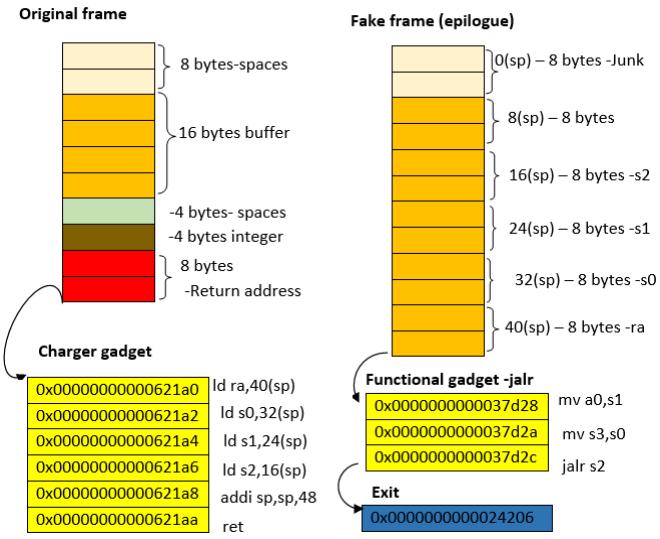


Figure 4: **Chaining gadgets to invoke a functional gadget**

4.3.3 Chained sequence to invoke a loop

ROP might not always end safely. The chained example in 4.3.2 can be transferred into a loop instead of *exit* if a gadget that does not include an instruction to overwrite the previous *ra* is inserted somewhere in the chain. When the functional gadget *getpid* (figure 5) was attached just after the *jalr* gadget, the program crashes because the *ra* returns back into the next instruction after *0x0000000000037d2c*. On the other hand, when the *jalr* gadget was taken out of the chain and the *getpid* gadget was chained directly to the charger, the return instruction directs execution back to the top of the *getpid* gadget and continues in a loop as shown in figure 6. This was observed in gdb but when the ROP chain was passed into the target at the command line, the execution only appeared to the user as if it was hanging. The execution did not crash but the process was not ending normally, resulting in a DoS.

```
00000000000362f4 <__getpid>:
   362f4:       0ac00893                li      a7,172
   362f8:       00000073                ecall
   362fc:       8082                    ret
```

Figure 5: **Extracted gadget *getpid***

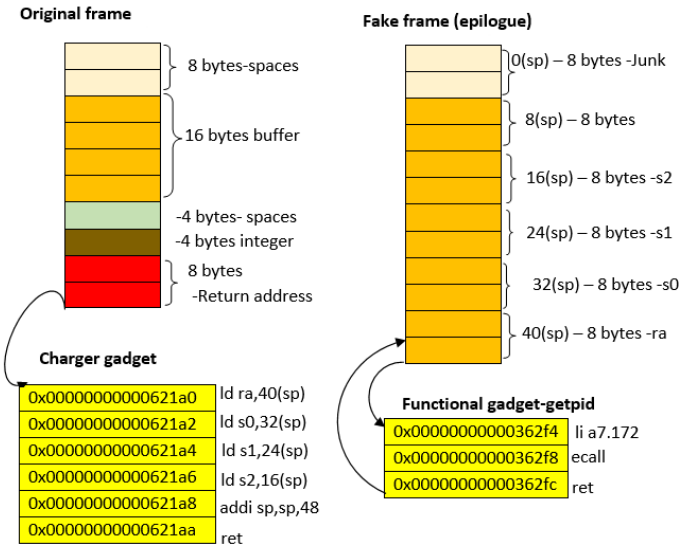


Figure 6: **Chaining gadgets to invoke a loop**

5. Discussion

The three categories of gadgets examined have shown different possible outcomes of ROP attacks based on the types of gadgets, the behaviours of each of the gadget, and the order in which the gadgets are chained. It could result in a full-blown exploit, a crash, or a loop. As for the *ret* gadgets, the expected operations are to pull address from the *ra* and jump into the address that was pulled to continue execution. This is what usually entails for return instructions but the peculiarity in RISC-V is that *ra* is a caller saved register. This means that upon each call to a subroutine, the return address is stored in the *ra* register. For this reason, the ROP chain needs to be crafted such that the intended return address for each subroutine is specified as demonstrated in figures 3 and 4. From the example in figure 6, it is highlighted that if the gadget does not include instructions to overwrite the value of *ra* then the execution would return to the value which has remained in the *ra* forming a loop in the execution. The same outcome was observed for *getgid* samples and could be for any gadget that does not enable the ROP chain to supply a new *ra*. The *getpid* and *getgid* gadgets exist in glibc and if used in this manner could become dangerous for the target.

None of the outcomes observed in section 4 is desirable on the path of the user. A loop in execution with no feedback to the user, is dangerous and could linger before it is identified. For the attacker, if they intended to achieve further exploits, passing execution to such gadgets would become an entrapment that truncates further execution of the ROP chain. The main finding here is that gadgets have peculiar behaviour based on the type of instruction at the end, the possibility of overwriting the *ra*, and the order in which the gadgets are chained. This

could be used in predicting the outcome of a ROP attack. Since the sample gadget used to cause a loop in execution here was picked from glibc, it is evident that this understanding could cut across other programs in the Linux environment and can be used broadly as information for training ML models in recognising possible ROP attacks.

Exploitation via ROP is still a challenge in software securities. Possible mitigation for the DoS demonstrated here is demonstrated by Oyinloye et al., (2022) using a watchdog monitoring process. The watchdog monitoring is based on timing-out depending on the log taken from the previous function epilogue. According to Oyinloye et al., the watchdog monitoring system can identify a looping ROP and stop or prevent DoS. A full exploit which would normally end undetected can also be stopped particularly if it extends reasonably for the watchdog to trace the time lapse after the last legitimate function. From the study, it is observed that ROP chains in RISC-V would require to be longer for any reasonable exploit to be achieved. It is considered that such long chain might be traceable by the watchdog monitoring process before a full exploit is achieved.

6. Conclusion and future works

ROP on RISC-V is not straightforward. The behaviour of gadgets in a RISC-V based process would result in different outcomes. While there are existing mitigations, it is important to consider the behaviour of gadget chains particularly when they access glibc. More vulnerable programs can be evaluated in order to obtain valuable information which could be harnessed in building algorithms for ML towards intrusion detection. More samples of exploitation might also reveal further interesting details. Future works would compare gadgets extracted from various programs that are known to be vulnerable to ROP including remotely executed programs. These programs can be compiled on RISC-V and X86, and then exploited with various sequences of ROP chains to observe the outcomes. A comparison of the behaviours of similar gadgets on these two platforms would be useful in providing a cross platform information for ML/DL models or other Control Flow Integrity concepts for combatting ROP. Extracts from the new gadgets will also be analysed to identify the reliability of the gadgets for ROP attacks. This would involve considerations under different compilers as well as different system architectures.

References

Bletsch, T., Jiang, X., Freeh, V. W., and Liang, Z., 2011. Jump-oriented programming: a new class of code-reuse attack. *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security,* March.pp. 30-40.

Carlini, N. and Wagner, D., 2014. ROP is still dangerous: breaking modern defenses. *SEC'14: Proceedings of the 23rd USENIX conference on Security Symposium,* August.p. 385–399.

Checkoway, S., Davi, L., Dmitrienko, A., Sadeghi, A-R., Shacham, H., and Winandy, M., 2010. Return-oriented programming without returns. *CCS '10: Proceedings of the 17th ACM conference on Computer and communications security,* October.pp. 559-572.

Deac, B., 2022. *InfoSec Write-ups.* [Online] Available at: https://infosecwriteups.com/return-oriented-programming-on-risc-v-part-1-dd9817b52d2b[Accessed 25 06 2022].

Gu, G. and Shacham H., 2020. No RISC No Reward:Return-Oriented Programming on RISC-V. 29 July.

IIT Madras, 2020. *IIT Madras, Indian Institute of Technology Madras.* [Online] Available at: https://www.iitm.ac.in/happenings/press-releases-and-coverages/iit-madras-develops-and-boots-moushik-microprocessor-iot[Accessed 08 July 2022].

Jaloyan, G-A., Markantonakis, K., Akram, R. N., Robin, D., Mayes, K., and Naccache, D., 2020. Return-Oriented Programming on RISC-V. *ASIA CCS '20: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security,* October.p. 471–480.

Koranek, D.F., Graham, S. R., Borghetti, B.J., and Henry, W. C., 2022. Identification of Return-Oriented Programming Attacks Using RISC-V Instruction Trace Data. *IEEE Access,* Volume 10, pp. 45347-45364.

Microsoft Coporation, 2018. *Microsoft Documentation.* [Online] Available at: https://docs.microsoft.com/en-us/windows/win32/memory/data-execution-prevention[Accessed 22 April 2020].

Oyinloye, T., Speakman, L., Eze, T., and O'Mahony, L., 2022. Watchdog Monitoring for Detecting and Handling of Control Flow Hijack on RISC-V-based Binaries. *International Journal of Advanced Computer Science and Applications(IJACSA),* 13(8).

Pax Team, 2003. *ASLR documentation.* [Online] Available at: https://pax.grsecurity.net/docs/aslr.txt[Accessed 22 April 2020].

RISC-V Community News, 2022. *RISC-V.org/blog.* [Online] Available at:https://riscv.org/blog/2022/07/deepcomputing-and-xcalibyte-open-pre-orders-for-first-native-risc-v-development-laptop-quantities-limited-xcalibyte-and-deep-computing/[Accessed 31 August 2022].

Shacham, H., 2007. The geometry of innocent flesh on the bone: return-into-libc without function calls (on the x86). *ACM conference on Computer and communications security,* pp. 552-561.

Basic Elements of Cyber Security for a Smart Terminal Process

Jouni Pöyhönen, Jussi Simola, Martti Lehto
University of Jyväskylä, Jyväskylä, Finland

jouni.a.poyhonen@jyu.fi
Jussi.hm.simola@jyu.fi
martti.j.lehto@jyu.fi

Abstract: Global maritime transportation and logistics systems are essential parts of critical infrastructures in every society, and a crucial part of maritime logistics processes are seaports. Digitalization helps improve the efficiency of terminal systems in the processes of these ports. In Finland this development is going on and it is called SMARTER research program. In the best cases, digitalization can also promote the reduction of emissions by optimizing port operations and enhancing cargo and people flows while improving the experience for all stakeholders. The improvement of port processes relies on the development of Information and Communication Technology (ICT) and as well as on Industrial Control Systems (ICS) or Operation Technologies (OT). At the same time, the cyber security aspects of maritime logistics also need to be addressed. In Finland, the SMARTER research program has been established to create port services by using, among other tools, Industry 4.0 solutions. As critical systems become more complicated in terms of users, processes and technology, the entire port infrastructure becomes a complex system of systems environment characterized by a conglomeration of interconnected networks and dependencies. ICT systems are significant parts of the operations and core processes, and are related to the administration and management of information in the network. The components of process levels include ICS/OT systems as well. This paper presents findings related to the SMARTER research goal on cyber security, which is a comprehensive cyber security architecture for port services at the system level. The paper emphasizes the importance of system description and recognizing basic system elements and their description in the first phase of the research process. The descriptions of these elements are needed to answer the following question at the beginning of the research: "What are the basic elements of cyber security for a smart terminal process?" The main elements identified in this paper are activities, the recognition of every stakeholder and the relationships between them, security dimensions, security capabilities, and system views of organizational criteria for cyber security in the SMARTER system. The solution can be called the Smart Port Cyber Security Management System (PortCSMS).

Key words: Maritime Logistics, Smart Port, Cyber Security Management, SMARTER

1. Introduction

The ENISA port cybersecurity report (2019) emphasizes the importance of maritime transport systems for the economy of the European Union. The report refers to activity that encompasses more than 1,200 seaports within the European Union, each of them with different organization, interests, challenges, and activities. On future development, the report states the following: "The global digitalization trend and recent policies and regulations require ports to face new challenges with regards to Information and Communication Technology (ICT). Ports tend to rely more on technologies to be more competitive, comply with some standards and policies and optimize operations" (ENISA, 2019).

International and national maritime transportation systems are essential parts of critical global infrastructures. Digitalization and increased levels of autonomy in logistics transport chains are expected to take leaps forward in the coming years. This development can help create safer, more efficient, sustainable, and reliable port services and operations that will have a central role in future transport logistics and supply chains. At the same time, as stated in the ENISA report: "This brings new stakes and challenges in the area of cybersecurity, both in the Information Technologies (IT) and Operation Technologies (OT) worlds" (ENISA, 2019). Digitalization makes it possible to create smart ports and terminals by utilizing the latest Industry 4.0 technology. A well-built digital port infrastructure is essential for optimizing operations and planning for future investment and maintenance needs.

In Finland, the Smart Terminals (SMARTER) research project, which runs to the end of February 2023, addresses the digitalization development of ports. The research work for port solutions benefits maritime transportation in many ways. For example, it will enable ports to reduce emissions by optimizing operations and improving cargo and people flows while improving the expertise 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, 2020b).

The digitalization of ports means the development of solutions for Information and Communication Technology (ICT), Information Technology (IT) and Industrial Control System (ICS) or Operation Technologies (OT). The maritime ports consist of digitalized system of systems where responses to system-level threats need to be coordinated as hybrid responses, hence the need for a system-of-systems–level research view. Such a view is necessary to address the relevant cyber safety aspects of the overall maritime solutions. In any cyber environment, trustable information networks are crucial. In addition, within operating environments where cyber security risks are continuously being highlighted by the threatening scenarios posed by the digital world, the usability, reliability, and integrity of systems data needs to be high. A modern society depends entirely on a cyber environment that provides dynamic services.

This paper follows our first SMARTER paper – "Emerging cyber risk challenges in maritime transportation" (Simola & Pöyhönen, 2022) – and leverages the research approach for the investigation of cyber security aspects at the system level in the study case. This is the next step in our research process and provides a description of a port environment and its basic elements. The paper highlights the importance of system description of the port process and its elements at the beginning of the research program. It also addresses our initial research question: "What are the basic elements of cyber security for a smart terminal process?" Once this question is addressed, the research will be able to specify the comprehensive cyber security aspects of architecture, such as threat analysis, risks assessment and cyber security measures, for the SMARTER project. This paper is one of the outcomes of the project's final report.

2. Smart Terminals (SMARTER) research project in Finland

In Finland, the Sea4Value / Fairway (S4VF) research program (DIMECC, 2020a) has moved on to its second phase. The SMARTER project enlarges the scope of DIMECC Sea4Value program to harbors and ports by developing digitalization and solutions that benefit RoPax and RoRo harbors. The project goals are the reduction of emissions by optimizing harbor operations and improving cargo and people flows while improving the experience for all stakeholders (DIMECC, 2020b). The first stage of the program concerns automated remote pilotage fairway navigation. The SMARTER project can 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, 2020b):

- Reduce emissions by optimizing port logistics.
- Enable exceptional flow and experience for passengers and cargo.

The structure of the project is built around three use cases (Figure 1): ship turnaround, truck traffic, and passenger flow. These use cases are designed to support one another and there is linkage between them. The applied research work is organized into five work packages, each led by a responsible team of researchers. Cyber security research actions are included in Work Package 4 (WP4). (DIMECC, 2020b)

The research items in Work Package 3 (WP3) on data collection and management are related to sensors, data processing, and connectivity to support data collection from diverse use cases. WP4 also provides cyber security research for comprehensive security management in the port environment. Cyber security will be integrated into the solution because information connections have high-level risks for attack vectors of adversarial actors. (DIMECC, 2020b)

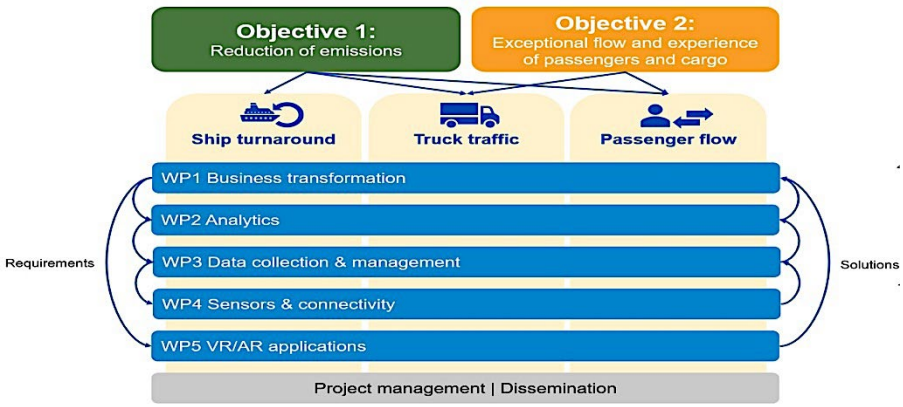


Figure 1: Project structure

The case of a port in a maritime transportation system includes processes such as ship approach from the open sea via a fairway to berthing at a pier, as well as port services, port logistics, and connections to land transportation. It is also obvious that the entity requires cooperation and communication with different stakeholders of the process elements. In all cases of port processes, the information requirements and the amount of information needed are related to the reliability of safety and security services. Cyber security awareness and information should cover all process elements. Figure 2 presents these processes. (Simola & Pöyhönen, 2022)

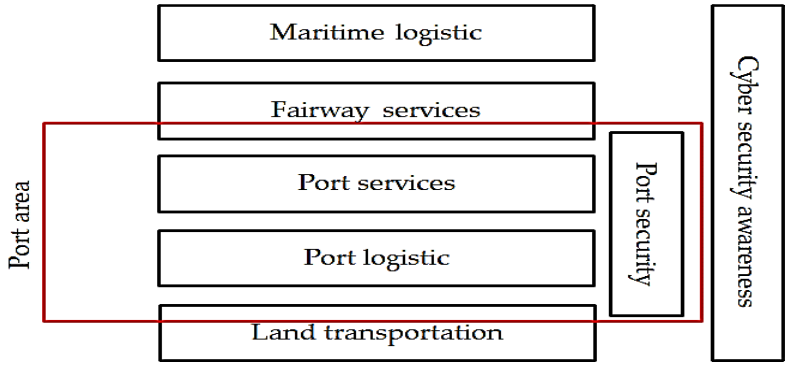


Figure 2: Port Processes

3. Industry 4.0 in the port processes

PwC's 2016 Global Industry 4.0 Survey mentions that the move to Industry 4.0 "is being driven by digitization and integration of vertical and horizontal value chains, digitization of product and service offerings and the development of new digital business models and customer access platforms" (PricewaterhouseCoopers, 2016). Since that time, an increasing number of papers have addressed, for instance, Industry 4.0 ontology, related technologies and manufacturing solutions. Moreover, ports and terminals in the 2010s have been seen to evolve to a new stage of industrial evolution, one characterized by their digital transformation and alignment with Industry 4.0 practices (de la Peña Zarzueloa, Soeanea & Bermúdez, 2020).

SMARTER acknowledges the need for new technologies and solutions that are necessary to tackle the challenges set by the use cases (DIMECC, 2020b). This set of technologies and solutions can be identified by the terminology of Industry 4.0. The development of smart port processes in the work packages, from control management and data collection to sensors and connectivity, needs to include definitions from big data, data lake, data analytics, information fusion, AI, 5G, IoT, edge processing and so on. From the Industry 4.0 point of view, this terminology is quite the same, so it is logical to speak about Industry 4.0 technology development as a general term related to smart terminal research goals. The future of smart ports will lead to increasing digital entry points to networks and thus connections between ICT systems and ICS or OT systems. The article "Cybersecurity Risks and Automated Maritime Container Terminals in the Age of 4IR (Fourth Industrial Revolution)" (Beaumont, 2018) presents a case study of automated maritime Container Terminals (CT) and demonstrates that the risks derived from the use of technology associated with the Fourth Industrial Revolution (4IR, referred to here as Industry 4.0) are both real and dangerous. He argues "that many CT operations are likely to be exposed to significant cyber-based risks and that this exposure will increase with the roll-out of further 4IR technologies unless appropriate control measures are implemented" (Beaumont, 2018).

Digital transformation also changes our thinking about separated systems. According to Accenture (2022), one example of digital transformation and converging IT/OT systems has been introduced in the mining industry. A copper mining company combined separated IT and OT systems under a joint governance structure where the management of IT and OT together were unified into one centralized technology organization.

4. Smart Port Cyber Security Management System

In their paper, "The Maritime Security Management System: Perceptions of the International Shipping Community" Thai and Grewal (2007) emphasize the importance of identifying key shore-based and near-shore activities associated with maritime security management. In their paper "Basic Elements of Cyber Security for an Automated Remote Piloting Fairway System" Pöyhönen, Kovanen and Lehto (2021) presents the findings of

cyber security research on the Sea4Value / Fairway (S4VF) project in the same way as in the fairway remote pilotage environment. Based on these papers there are similar needs to identify key elements associated with operations and processes used in the SMARTER research areas. These are listed on the heading level as follows: activities, stakeholders, organizational relationships, security dimensions, security capabilities, and criteria. Figure 3 illustrates the cyber security elements of this study. To cover all these, the result can be called Smart Port, Cyber Security Management System (PortCSMS). The cyber security elements are explained after the figure.

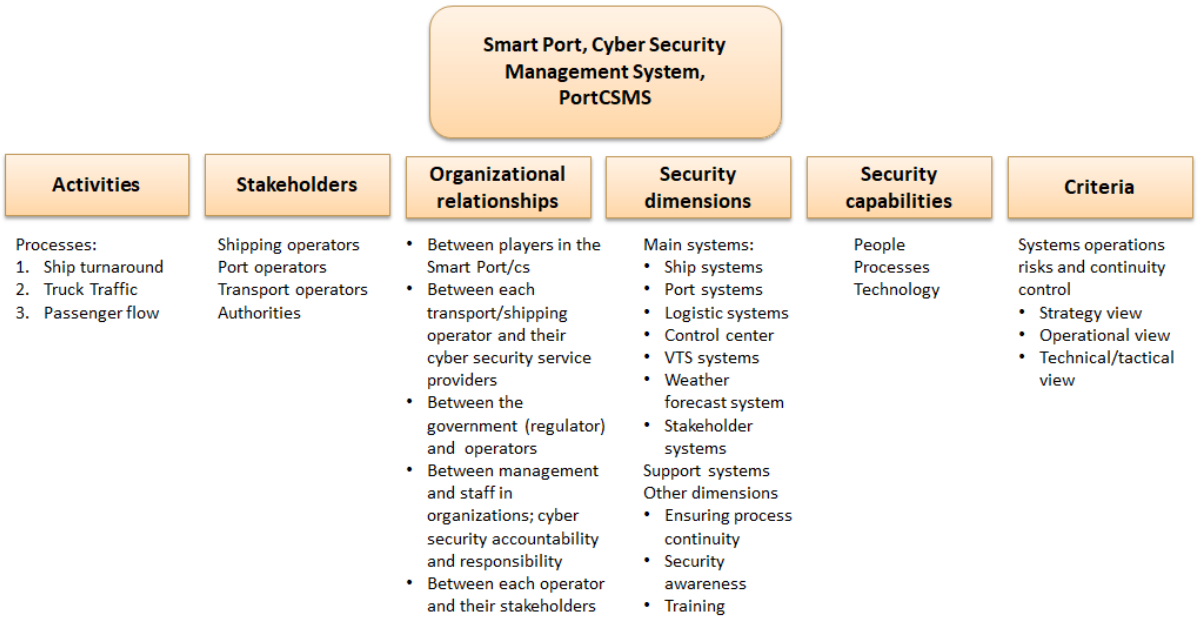


Figure 3: Elements of smart port cyber security management, PortCSMS.

4.1 Activities

The activities in this port research are ship turnaround, truck traffic, and passenger processes. They also comprise the use cases of the research. "The ship turnaround use case focuses on optimizing the various processes associated with ship visits in harbor. The truck traffic use case focuses on developing a predictive real-time traffic situational awareness model to optimize the cargo logistics to and from the port. The passenger flow use case improves the passenger flow on RoPax ports by developing models and solutions to support collective transportation, ride sharing and other mechanisms to relieve the congestion at ports." (DIMECC, 2020b)

All ports play an important role in maritime business processes and, by extension, in global logistics. Ports are the place where maritime logistics connect to land logistics, making them one of the main gateways in the international trade system. The strategic positioning of ports has significantly changed in the last two decades. The standards of their operations and service management have increased because of the global competitive landscape. In addition, the ongoing digitalization of ports is going to be a major driver in their development in the near future. Smart solutions for ports enable identifying new growth areas as well as changing the revenue mix, logistics operations, and port services. In this environment the largest challenge can be tackled by using the latest solutions for digitalization. (DIMECC, 2020b)

The ENISA Threat Landscape Report 2016 emphasizes all elements covered within an attack on a business process. It means that not all of the artefacts/components used are IT related. There are steps and procedures used within an attack that are performed by just having knowledge or information about the details of the business process at stake. Business-related issues are key areas, both in the planning of an attack by an adversary and in analyzing an incident by a defender of activities: "Asset exposure can be grouped based on a business process or an asset owner" (ENISA, 2017).

4.2 Stakeholders

A modern seaport can include dozens of stakeholders interacting to run the port processes. In a stakeholder organization, the cyber security capabilities are based on the expertise of people, trustable processes and security technology in products and services. At the strategy level this means that company management includes organization leaders who consider cyber security issues as a strategic goal, a published cyber security policy, and risk-based management as part of overall cyber security and business activity (Pöyhönen & Lehto, 2017).

At the operational level, the strength of the stakeholder cyber security approach may be based on the use of the best partners in outsourcing, clustering, public–private partnerships (PPP) and international cooperation. A good reputation among stakeholders is needed in this approach. Of course, stakeholders should work without conflict between a company's business activity and the national supply security requirements (resourcing) for critical infrastructure. (Pöyhönen & Lehto, 2017)

The technological-tactical level implementation focus should be on staff competence, cyber and information security products and services. In the best cases, the strength of companies is based on ensuring the continuity of operations through training, planning exercises, and preparedness plans. Real-time situation awareness of ICT assets and ICS/OT systems is mandatory to have at least in technological-tactical level. (Pöyhönen & Lehto, 2017)

4.3 Organizational relationships

Port processes and services include the relationships among organizations within the harbor as well as those outside of it. It is possible to have companies' common situation awareness and to have the opportunity to learn about threats often directly from the operating network or partners and the use of announcements of the National Cyber Security Centre Finland. On the other hand, overall situation awareness is often based on scattered data, and obtaining situation awareness of the entire operating network is useful but could be at the same time challenging. (Pöyhönen & Lehto, 2017)

A common port facility cyber security plan (CSP) is needed to manage the risks of a security incident. It is intended that wherever appropriate, the CSP will build upon the existing port security plan (PSP). (IET, 2020)

In general, a CSP covers the strategy, policy, procedures, and technologies of an organization. The port facility security plan is intended to ensure port processes, stakeholders, and relationships between them and as well as the port systems. Port stakeholders can leverage expert services as a strength in different audits and in solving problem situations also by utilizing best practices and national research programs (Pöyhönen & Lehto, 2017).

4.4 Security dimensions

A. Main systems

A block diagram makes a description of the cyber security main dimensions for a system-of-systems approach in port operations. Figure 4 illustrates these blocks (Simola & Pöyhönen, 2022). The figure shows the functional sides of the port that are needed in order to understand the basic architecture of its cyber security management.

Communication and functional relationships between the blocks in the port systems are one of the key features of its operational processes. They are essential elements to be studied in this research process.

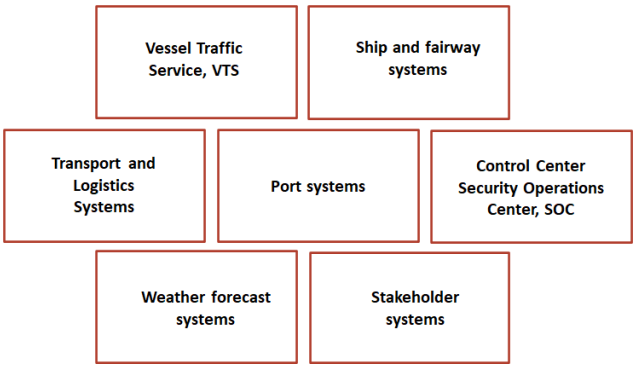


Figure 4: Block diagram for Port Operations

Figure 4 specifies the key cyber security subsystems in a block diagram created as part of the SMARTER project. Its foundation is the main port process, including transport and logistics systems, port systems and control center systems. The port operations also need support processes such as vessel traffic service (VTS) and weather forecast services. Ship and fairway systems and stakeholder operations are also essential information sources and support processes. The main process elements of the port are outlined in the middle row and the support processes are in the top and bottom rows in the figure. The complete system configuration is a complex system-of-systems environment. All of these systems could face cyber threats in many ways.

B. Support systems

Based on our previous articles (Pöyhönen & Lehto, 2017; Pöyhönen, 2022), the concept of national critical infrastructure can be simplified in accordance with three essential system layers. At the base layer is the electricity network, above that is the information network layer and above these are services. Each layer in system-of-systems thinking can assume its own strategic role and identify its operation as part of an entity whose other parts depend on a reliable functioning of these layers. This also facilitates the identification of cyber dependencies within the layers so that they can be secured with the most efficient and practical measures. In this port research project, electrical power systems and networks are the support systems for the port services. Due to known threat scenarios, these are also security dimensions that should be taken into account as part of any cyber security assessment.

C. Other dimensions

Ensuring process continuity. Process continuity is related to the systems of critical infrastructures. In 2016, the European Parliament adopted the Directive on security of network and information systems (the NIS Directive), with the aim of bringing the cybersecurity capabilities for the essential services of critical infrastructures to the same level of development in all EU Member States. Doing so means that exchanges of actions, information and cooperation should be efficient within and between the organizations of essential services, including at the cross-border level. Operational risk is a crucial part of the prudent regulation and supervision in the sectors of critical infrastructures. It covers all operations including the security, integrity and resilience of network and information systems. The core idea of the NIS directive is that the relevant service operators must ensure business continuity in the case of adverse information security disruptions and report any substantial information security breaches to authorities. According to ANNEX-II of the directive, the transport sector is one of the sectors to be covered by the regulations of the NIS Directive. (EU, 2016; Vähäkainu, Lehto, & Kariluoto, 2022)

Smart ports, as part of critical infrastructure, are facing various cyber threats that may lead to the appearance of events that will cause the disruption or failure of services processes. It is useful to minimize the impact of cyber disruptions and thus ensure the continuity of services by strengthening system resilience. Resilience can be seen as an organization's capacity and capability to achieve its purposes in both predictable and unpredictable situations or under continuous stress caused by cyber security threats (Pöyhönen et al., 2018). The resilience metrics framework proposed by Linkov et al. (2013a) is utilized by applying resilience measures to an organization's operational processes. The framework combines the four stages of a system (plan/prepare, absorb, recover, and adapt) with the four domains of a system (physical, information, cognitive, and social). Later Linkov et al. (2013b) have applied their model further to cyber systems.

Security awareness. Efficient cyber security management in general and, in this case, for smart ports requires close collaboration among management, situational awareness (SA), and communication. In that sense, good management requires the casting of different port operators and their decision-making ability and the building of real-time situation awareness of ICT and ICS/OT assets as well as a common cyber environment (situational understanding, evaluation of situational development). The capabilities of crisis communication, information sharing, and supporting technical solutions, business continuity management, and cooperation are also needed for the good practice of security management. Such management is a combination of the actions of all the organization's decision-making levels (strategic, operational, and technical/tactical) and utilizes the national and international strengths of information sharing. (Pöyhönen, et al. 2021)

The security operation center (SOC) is a centralized cyber security feature of an organization employing people, processes, and technology to continuously monitor an organization's security situation. It has the ability to prevent, detect, analyze, and respond to cybersecurity incidents. In the case of ports, it is recommended as a good practice: "The SOC acts as a centralised unit dealing with security issues that affect a port/port facility, including those relating to cyber security, and may form part of an operations centre supervising the port,

controlling access and managing business continuity and disaster recovery activities" (Boyes, Isbell, & Luck, 2020).

SA can be implemented based on SOC information and national as well as international information sources. Ensley (1995) has developed an SA model when working in the service of United States Air Force. According to her, the core of SA consists of three basic levels: detection (Level 1), situational understanding (Level 2), and its impact assessment towards the future (Level 3) (Ensley, 1995). A comprehensive SA of port processes provides a foundation for conclusions and the subsequent decision-making in cyber security incidents.

Training. The continuous improvement of activities related to the development of staff competence enhance an organization's capability. Taking the staff into account at all organizational levels, as well as focusing on competence and the possibilities it opens to fully influence in the organization, develops the overall operations of the organization (Finnish Standards Association SFS, 2016). Cyber security training as well as other development of staff competence enhance an organization's capability to proactively prevent disturbances and tolerate potential changes in process operation such disturbances might cause.

The maritime sector is increasingly at risk of cyber-attacks due to advances that are already in the process of being implemented, such as port services that utilize new technologies. For that reason, the level of knowledge and training on cyber security and its interaction with the marine ecosystem must be continuously maintained. Recent research results on the topic show a lack of general knowledge in the field of maritime cyber security. Therefore, it is necessary to establish cyber security training as one of the security dimensions in the maritime field. (Alcaidea & Llave, 2019)

4.5 Security capabilities

Different cyber security elements are related to various components, including people, processes, and technology. In their framework, Thai and Grewal emphasize people, processes, and systems/technology as one of the most important elements. The human factor has always been seen as a key factor in any security management system. Concurrently, the importance of people and several combinations of this element, such as people and communication, people and processes, and people and systems/technology, are highly visible. (Thai & Grewal, 2007)

Jacobs, von Solms, and Grobler (2016) state that "the governance documents of an organization typically [prescribe] sets of controls to be implemented, such as technical controls, administrative controls, and physical controls." In many cases these documents describe very specific capabilities that are needed and have to be developed in securing operational continuity in the cyber domain. Capabilities consisting of "people, processes and technology" are key elements in achieving secure outcomes or effects.

4.6 Criteria

An organization's cyber security operations require comprehensive awareness on the system level. Apropriate awareness thus supports cyber risk management and, more extensively, the evaluation of an organization's whole cyber capability. By integrating an organization's three main decision-making levels (strategy, operational, and technology/tactical) into the structure of its cyber operating environment, it is possible obtain a holistic system view of an organization's cyber security tasks. 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 can be considered a framework for evaluating cyber security management. (Pöyhönen & Lehto, 2020).

For the goals of cyber security in a smart port environment, the risk management and continuity enhancement of system operations establish 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.

5. Discussion

ENISA report (2019) "Port Cybersecurity" identifies good practices for cybersecurity in the maritime sector taking care undergoing digital transformation in ports operations and processes. In order to meet emerging challenges is recommended to optimise existing processes and introduce new capabilities, such as automation and real-time monitoring of operations. Digital transformation trends and policies and regulations require ports to face new challenges with regards to the Information and Communication Technology (ICT) and as well as on Industrial

Control Systems (ICS) or Operation Technologies (OT) worlds. The report underlines need to have a high-level reference model based on the research and information of port structure. Its objectives are to list, from a high-level perspective, the main port systems, data flows and interactions with external systems.

The Institution of Engineering and Technology, IET, (2020) in "Good Practice Guide, Cyber Security for Ports and Port Systems", says that "a port is a complex cyber environment that encompasses both land and waterside activities and systems". The loss of cyber security in one section, or more, of port assets has the potential to impact upon efficiency at the port operations, safely of operations or the health and safety of staff and other people.

The Directive 2016/1148 (NIS Directive) is an EU-wide cybersecurity legislation. It covers the maritime sector as one of the essential services under Union legal acts in people, processes and technologies of ports. It recommends also that member States should take into account existing and future international codes and guidelines developed in particular by the International Maritime Organisation. (EU, 2016)

IMO (2017) "Guidelines on Maritime Cyber risk Management" provides high-level recommendations for maritime cyber risk management as an essential measure to ensure the continuity of operations and processes in maritime.

The aim of this paper "Basic Elements of Cyber Security for a Smart Terminal Process" is to give the description of the smart port environments at the system of systems level and thus makes it easier to perceive the cyber security requirements described above.

6. Conclusion

This paper provides a holistic research approach for the investigation of the cyber security of the system elements of smart terminal processes in the development of port operations. It includes system-of-systems thinking and it emphasizes the importation of system description at the beginning of the research project. The findings of the study are related to the basic elements of port operations. These are port services for ship turnaround, truck traffic, and passenger processes. SMARTER research activities are related to digitalization survey with the use of new technologies in order to achieve the main objectives, which are the reduction of emissions by optimizing port logistics and enabling exceptional flow and experience for passengers and cargo.

Comprehensive cyber security research needs to understand the basic system elements of smart terminals as well as the content of those elements. The basic elements have been identified and the result can be called PortCSMS. According to PortCSMS, these elements are as follows: activities, stakeholders, organizational relationships, security dimensions, security capabilities, and criteria. The contents of these are also listed. In addition to these elements, a research framework is needed to address the questions. The next steps are to study the comprehensive cyber security architecture as part of SMARTER program.

Future research on cyber security in the SMARTER project will now be based on what the basic elements of the port are determined to be. After this, it will be possible to work towards answering the main research question of the project: How can a comprehensive cyber security architecture for smart port processes be developed?

References

Accenture. 2022. Digital Transformation through IT/OT convergence. Accessed: 26/9/2022. https://www.accenture.com/fi-en/case-studies/natural-resources/digital-transformation-through-it-ot-convergence

Alcaidea, J. I. & Llave, R.G., 2019. Critical infrastructures cybersecurity and the maritime sector. AIIT 2nd International Congress on Transport Infrastructure and Systems in a changing world (TIS ROMA 2019), 23rd-24th September 2019, Rome, Italy. ScienceDirect. Transportation Research Procedia 45 (2020) p. 547–554.

Beaumont, P., 2018. Cybersecurity Risks and Automated Maritime Container Terminals in the Age of 4IR [Fourth Industrial Revolution], Chapter in 'Handbook of Research on Information and Cyber Security in the Fourth Industrial Revolution', pp. 497–516, https://doi.org/10.4018/978-1-5225-4763-1.ch017

Boyes, H., Isbell, R: & Luck, A., 2020. Good Practice Guide Cyber Security for Ports and Port Systems. Institution of Engineering and Technology, London, United Kingdom.

de la Peña Zarzueloa, I., Soeanea M. J. F. & Bermúdez, B. L., 2020. Industry 4.0 in the port and maritime industry: A literature review. Journal of Industrial Information Integration 20 (2020) 100173.

DIMECC Oy, 2020a. SEA FOR VALUE (S4V). 12.2.2020. https://www.dimecc.com/dimecc-services/s4v/

DIMECC Oy, 2020b. DIMECC Sea4Value/Smart Terminals (SMARTER). Project proposal for One Sea – autonomous maritime ecosystem.

Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. Hum Factors Ergon Soc 37(1), p. 32–64.

ENISA, 2017. Threat Landscape Report 2016, 15 Top Cyber-Threats and Trends.

ENISA, 2019. PORT CYBERSECURITY. Good practices for cybersecurity in the maritime sector. NOVEMBER 2019. https://www.enisa.europa.eu/publications/port-cybersecurity-good-practices-for-cybersecurity-in-the-maritime-sector

EU, 2016. Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. The European Parliament and the Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016L1148

Finnish Standards Association SFS, 2016. Johdanto laadunhallinnan ISO 9000 -standardeihin. Available on 20th September 2022: slideplayer.fi/slide/11133323/

IET, 2020. Good Practice Guide. Cyber Security for Ports and Port Systems. The Institution of Engineering and Technology. UK.

IMO, 2017. GUIDELINES ON MARITIME CYBER RISK MANAGEMENT. MSC-FAL.1/Circ.3 5 July 2017. International Maritime Organization.

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, p. 51–61.

Linkov, I., Eisenberg, D., Bates, M., Chang, D., Convertino, M., Allen, J., Flynn, S. and Seager, T., 2013a. Measurable Resilience for Actionable Policy. Environmental Science & Technology.

Linkov, I., Eisenberg, D., Plourde, K., Seager, T., Allen J. and Kott, A., 2013b. Resilience metrics for cyber systems. Environment Systems and Decisions, 33(4), p. 471-476.

PricewaterhouseCoopersin (PwC), 2016. Industry 4.0: Building the digital enterprise. https://www.pwc.com/gx/en/industries/industrial-manufacturing/publications/assets/pwc-building-digital-enterprise.pdf

Pöyhönen, J. & Lehto, M., 2017. Cyber security creation as part of the management of an energy company. Proceedings of the 16th European Conference on Cyber Warfare and Security ECCWS2017. 29 - 30 June 2017, Dublin, Ireland. Published by Academic Conferences and Publishing International Limited. Reading. UK. p. 332-340.

Pöyhönen, J., Nuojua, V., Lehto, M. & Rajamäki, J., 2018. Application of Cyber Resilience Review to an Electricity Company. Proceedings of the 17th European Conference on Cyber Warfare and Security ECCWS2018, 28 – 29. June 2018, Oslo, Norway. Published by Academic Conferences and Publishing International Limited. Reading. UK. p. 380-389.

Pöyhönen, J., Rajamäki, J., Nuojua, V., & Lehto, M., 2021. Cyber Situational Awareness in Critical Infrastructure Organizations. In T. Tagarev, K. T. Atanassov, V. Kharchenko, & J. Kacprzyk (Eds.), Digital Transformation, Cyber Security and Resilience of Modern Societies (p. 161-178). Springer. Studies in Big Data, 84. https://doi.org/10.1007/978-3-030-65722-2_10xx

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 ICCWS2021, 25 - 26 February 2021, Tennessee, USA. Published by Academic Conferences and Publishing International Limited. Reading. UK. p. 299-308

Pöyhönen, J. & Lehto, M., 2020. Cyber security: Trust based architecture in the management of an organization security. Proceedings of the 18th European Conference on Cyber Warfare and Security ECCWS2020, 25-26 June 2020, Chester, UK. Published by Academic Conferences and Publishing International Limited. Reading. UK. p. 304-313

Pöyhönen, J., 2022. Cyber Security of an Electric Power System in Critical Infrastructure. Cyber Security, Critical Infrastructure Protection. Martti Lehto Pekka Neittaanmäki Editors. Springer. Computational Methods in Applied Sciences. Volume 56. Chapter 9. (p. 217-254). ISSN 1871-3033. ISBN 978-3-030-91292-5 ISBN 978-3-030-91293-2 (eBook) https://doi.org/10.1007/978-3-030-91293-2

Simola, J. & Pöyhönen, J., 2022. Emerging cyber risk challenges in maritime transportation. Proceedings of the 17th International Conference on Information Warfare and Security ICCWS2022. 17-18 March 2022, Albany, New York, USA. Published by Academic Conferences and Publishing International Limited. Reading. UK. p. 306-314.

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.

Vähäkainu, P., Lehto, M. & Kariluoto, A., 2022. Cyberattacks Against Critical Infrastructure Facilities and Corresponding Countermeasures. Cyber Security, Critical Infrastructure Protection. Martti Lehto Pekka Neittaanmäki Editors. Springer. Computational Methods in Applied Sciences. Volume 56. Chapter 11. (p. 255-291). ISSN 1871-3033. ISBN 978-3-030-91292-5 ISBN 978-3-030-91293-2 (eBook) https://doi.org/10.1007/978-3-030-91293-2

Anti-American Stance in Turkey: A Twitter Case Study

Gowri Prathap[1], Alex Korb[2], Luke Palmieri[1], Ekrem Kaya[1], Saltuk Karahan[2] and Hamdi Kavak[1*]

[1]George Mason University, Fairfax, VA, USA
[2]Old Dominion University, Norfolk, VA, USA
[*] Corresponding author

gprathap@gmu.edu
rkorb001@odu.edu
lpalmie@gmu.edu
ekaya4@gmu.edu
skarahan@odu.edu
hkavak@gmu.edu

Abstract: The availability of social media and biased actors exacerbated Anti-American and Anti-Western views to extremes. In this paper, we report our efforts in analyzing anti-American views on Twitter. We have collected over three years of Turkish tweets related to the US, translated them into English, and analyzed these tweets using various computational social science tools. We found that Turkish tweets related to the US are significantly negative, and emotions reflect disgust and anger. Furthermore, we found that the source of the negative views stems from political actors like Trump or Biden rather than general hatred. Our results shed light on potential policy plans and interventions.

Keywords: Targeted Sentiment Analysis, Emotion Analysis, Disinformation, Anti-Americanism

1. Introduction

As global Internet coverage and social media grow ever more widespread, the strategic potential of this cyberspace as a means of engaging rivals is beginning to be realized, such that cyberspace has been recognized as the fifth domain of force (Welch, 2011). Thus, it provides a platform for disseminating controlled or uncontrolled narratives from various actors. There is a vast amount of work on the information warfare aspect of cybersecurity, specifically on how public opinion is being shaped by external actors (Margolin, Hannak & Weber, 2018). Russia has been a major international actor in taking advantage of free speech and free information access in western countries and conducting cyber campaigns to gain an advantage (Narayanan, Howard, Kollanyi & Elswah, Mona, 2017). One of the well-known cases of involvement is the recent US elections, as confirmed by US intelligence agencies (Mueller III, 2019). Less-known Russian operations include the campaign against the opposition leader Alexei Navalny (Alieva & Carley, 2021), among many others (Beskow & Carley, 2020).

Furthermore, social media can also reflect the views of populations (Momtazi, 2012), people living in certain regions (Ghahramani, Galle, Ratti & Pilla, 2021), and even particular social groups (Schneider & Carpenter, 2019). Thus, we can mine social media data to (1) investigate how actors use it to gain an advantage and (2) understand people's stances against any topic. Computational social science tools help us tackle these tasks.

In this paper, we analyze anti-American views in Turkish Twitter messages published from 2019 to March 2022. We use various natural language processing techniques applied to these tweets to investigate the following questions.

- What are general sentiment and emotion trends in US-related tweets?
- How different are these trends in other Turkish tweets?
- What are the common targets of these trends?

1.1 Social Media and Disinformation in Turkey

Often referred to as "fake news", disinformation, or the production and targeted dissemination of misleading content to "generate profits, pursue political goals, or maliciously deceive", (Humprecht, Esser & Aelst, 2020) is just one of the ways of influencing cyberspace, and one that has come under increasing scrutiny by state and non-state actors. Disinformation is not limited to cyberspace, and it is possible for mass media news networks to produce disinformation on behalf of a foreign or domestic actor (Palmieri et al., 2022). Still, the global reach of social media platforms means that disinformation can be spread transnationally. Disinformation is not the same as misinformation, which is the unintentional spread of incorrect or misleading information, but misinformation can result from disinformation when spread by individuals who are disinformed (Humprecht,

Esser & Aelst, 2020). Disinformation is also not to be confused with influence operations, the leaking or dissemination of actual content to influence the domestic population of a target country, but influence operations are often included under the same umbrella as disinformation, and a prolonged influence campaign can include both. The means by which disinformation campaigns and influence operations are carried out are very similar: social media bots, troll armies, and astroturfing techniques are used on a massive scale (Kirdemir, 2020).

Disinformation campaigns are rampant in the Republic of Turkey, where President Recep Tayyip Erdoğan and the Justice and Development Party (AKP) take a direct hand in attempting to control social media narratives. In the aftermath of the 2013 Gezi Park protests, the AKP began to employ a troll army of roughly 6,000 users, nicknamed "AK Trolls", whose task is to promote pro-Erdoğan tweets, accounts, and narratives and harass opponents on Twitter (Akiş, 2022). Although Twitter suspended many of these accounts in 2020 (Benedictus, 2016), the AK trolls remain at large. In addition, a social media law proposed in August 2021 made spreading "fake news" punishable with up to five years in prison (Wilks, 2021); this, in combination with the social media censorship law of July 2020, demonstrates Erdogan's more formal attempts to control social media.

2. Data Collection

We collected Turkish tweets from Twitter using Twitter's Filtered stream API. We constructed two datasets; the first one is a dataset of USA-related Turkish tweets (collected by sending relevant keywords as inputs to the API). Some of the keywords are "abd", "amerika", "amerikan", "amerikanci", "coni", "biden", "trump", "sanders", "obama", "pentagon"; we also specified Turkish as the language in this filtered stream. The tweets are collected from January 2019 to March 2022. The second dataset consists of a sample of general Turkish tweets from September to December 2021. We collected a sample of general Turkish tweets to understand the general sentiment and emotions as a baseline and compared them with the Turkish tweets related to the USA. We translated these tweets to English using Google Translate's API.

When collecting and analyzing our data, we had to look for various ways to minimize bias. Our study wanted to look at Turkish sentiment towards the USA. However, there are Turkish people living across the world, so to minimize geographical bias, instead of collecting tweets that were tweeted in Turkey, we collected tweets in the Turkish language, to understand the sentiments of Turkish people across the world. Another bias is sentiment bias, which we investigated in greater detail. We collected tweet samples written in Turkish to understand the general sentiment and emotions as a baseline and compared them with the Turkish tweets related to the USA, to eliminate any sentiment bias. We also took all Turkish tweets with relevant keywords and not just a sample. We also avoided retweets

because certain tweets by government officials or politicians would get more traction if we counted retweets.

USA related Turkish tweets (before excluding retweets): From January 2019 to March 2022, there were a total of 47,629,360 tweets related to the USA containing at least one of the keywords. There were 10,347,907 tweets in 2019; 13,767,535 tweets in 2020; 17,380,699 tweets in 2021 and 6,133,219 from January to March 2022. The average number of daily Turkish tweets related to the USA is around 40,160.

USA related Turkish tweets (after excluding retweets): We excluded retweets from our analysis. From January 2019 to March 2022, there were a total of 13,941,788 unique Turkish tweets related to the USA. Of these, there were 2,939,517 tweets in 2019; 4,195,548 tweets in 2020; 4,853,341 tweets in 2021 and 1,953,382 tweets from January to March 2022. The average number of unique daily Turkish tweets related to the USA is around 11,755 tweets.

A sample of general Turkish tweets: The number of general Turkish tweets we collected from September 2021 to December 2021 is 4,226,680. Note that we missed data for a few days due to outages. Figure 1 shows the number of daily general Turkish tweets and Turkish tweets related to the US.

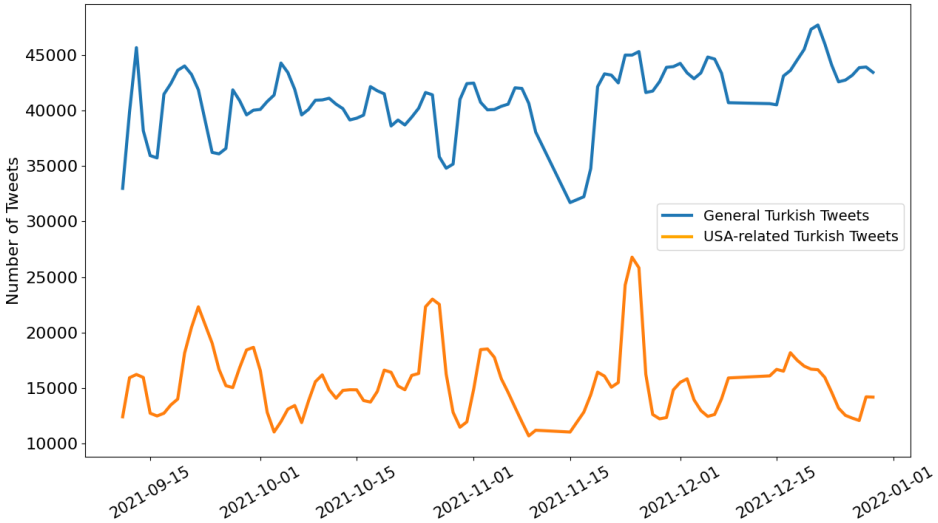


Figure 1. The size of general Turkish tweets vs. Turkish tweets related to USA

3. Methods

3.1 Classical Sentiment Analysis

Sentiment analysis of a text determines if it expresses emotions that are positive, negative, or neutral (Zhang, Wang & Liu, 2018). Each tweet is assigned a score that represents the entire text's sentiment level and polarity. If the sentiment score is less than 0, the text has negative polarity; if the score is equal to 0, the text is neutral; if the score is greater than 0, the text has a positive polarity. For conducting this classical sentiment analysis, we used Afinn (Nielsen, 2016), a Python on the English translations of the tweets. Afinn uses a wordlist-based approach to score sentences.

3.2 Targeted Sentiment Analysis (TSA)

Targeted Sentiment Analysis is more detailed than classical sentiment analysis in the sense that it identifies the sentiment and sentiment polarity of a text towards a particular target, such as a person or country. For example, a sentence like 'I like Mr. X, but I do not like Mr. Y' has two targets: Mr. X and Mr. Y. TSA can identify that the sentiment toward Mr. X is positive, while the sentiment toward Mr. Y is negative. The classical sentiment analysis for this sentence would have been neutral because one positive and one negative score of the same verb (i.e., "like") would cancel each other. TSA uses Named Entity Recognition (NER), which identifies targets such as an object or people in a text (Pei, Sun & Li, 2019). It allows us to identify the sentiments of the text towards different entities. TSA is also called aspect-based sentiment analysis in the literature (Pei, Sun & Li, 2019).

We performed TSA using the NewsSentiment package (Hamborg & Donnay, 2021). This package implements a model that can interact with multiple embeddings from, for instance, a pre-trained language model (e.g., Roberta (Liu et al., 2019) and other knowledge sources. NewsSentiment works on news and social media data and provides the probability that the sentiment is classified as positive, negative, or neutral towards a particular target. Since NewsSentiment requires targets up front, we performed TSA on the keywords 'Biden', 'Trump', 'The US', and 'The USA.'

3.3 Emotion Analysis

Emotion analysis of a text can identify emotions such as joy, anger, fear, and sadness (Zhang, Wang & Liu, 2018). We used Pysentimiento to conduct emotion analysis of the tweets (Pérez, Giudici & Luque, 2021). Pysentimiento is a transformer-based emotion analysis package based on the social language model BERTweet (Nguyen, Vu & Nguyen, 2020). As a result, Pysentimiento returns six particular emotions (joy, sadness, anger, surprise, disgust, and fear) and "other", all can have different scores adding up to 1.0.

3.4 Limitations

Our study has a few limitations. Some limitations were that sentiment and emotion analysis cannot identify sarcasm. In social media, sarcasm is prominent, and the algorithms are usually not able to identify this sentiment. Also, emotion analysis classifies the tweet's emotions into six that are too broad and need to be more specific. Another limitation is that Twitter data has a lot of noise. There is an abundance of trolls, bots, and spam tweets, which can skew our results. We also gave keywords to pull data, but this could have pulled in irrelevant data which happened to have that keyword.

4. Results

4.1 What are general sentiment and emotion trends in US-related Turkish tweets?

Figure 2 shows the daily average Afinn-based sentiment score of US-related translated tweets. There is a clear trend of highly negative average sentiment scores almost for the entire data collection timeframe. We notice a few positive scores here, the highest one recorded close to the 2020 US Presidential Elections.

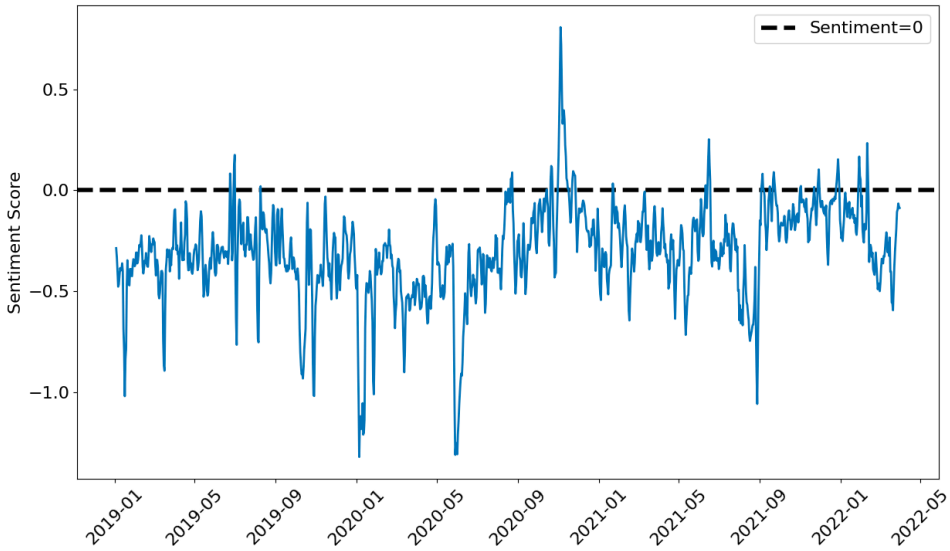


Figure 2: Afinn sentiment score

Figure 3 illustrates the distribution of different emotions over time. We observe the most prevalent emotion is *disgust*, followed by *anger*. We also observe some peaks, namely *anger* on Oct 7, 2019, *fear* from March 10-13 2020; and *disgust* on July 22, 2021 and from August 15-20, 2021. The anger was in response to Trump's threatening to "destroy and obliterate" the Turkish economy (Sonmez 2019) and possibly the Turkish invasion of Syria as well, since there are a lot of references to the US arming "terrorist organizations" (the YPG) (The Guardian, 2019). The fear peak could be in response to the US suspending travel to Europe, but some users were also getting angry over the New York Times and the BBC using pictures of mosques in Turkey to represent Europe (Allyn & Romo, 2020).

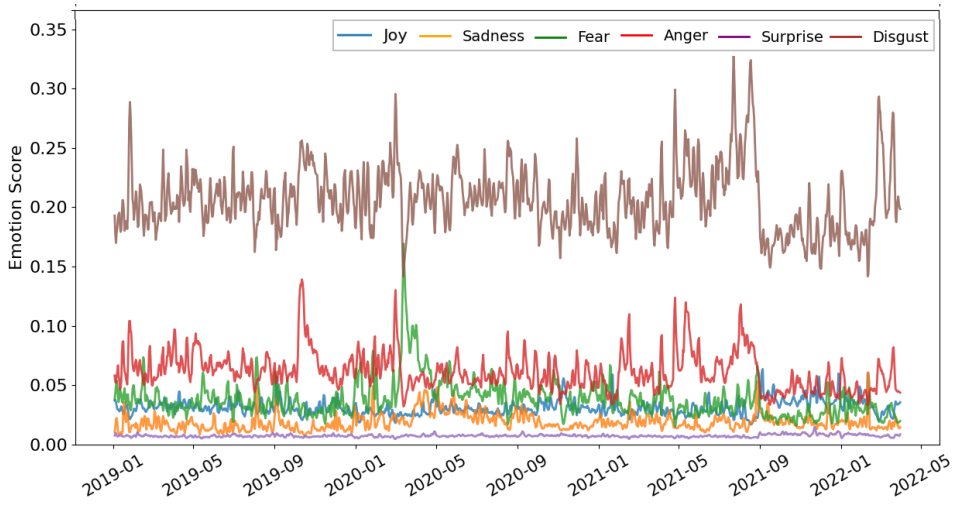


Figure 3: Emotion Analysis of Turkish tweets related to the USA

4.2 How different are these trends in other Turkish tweets?

Figure 4 shows the sentiment score for approximately three months in which we captured both USA-related and general Turkish tweets. We can observe that the general sentiment of Turkish tweets is positive, and the tweets

have a higher sentiment score in comparison to Turkish tweets related to the USA. This result suggests that the baseline sentiment score for USA-related tweets is negative, which reflects negative views against the US in Turkey (Global indicators database, 2021).

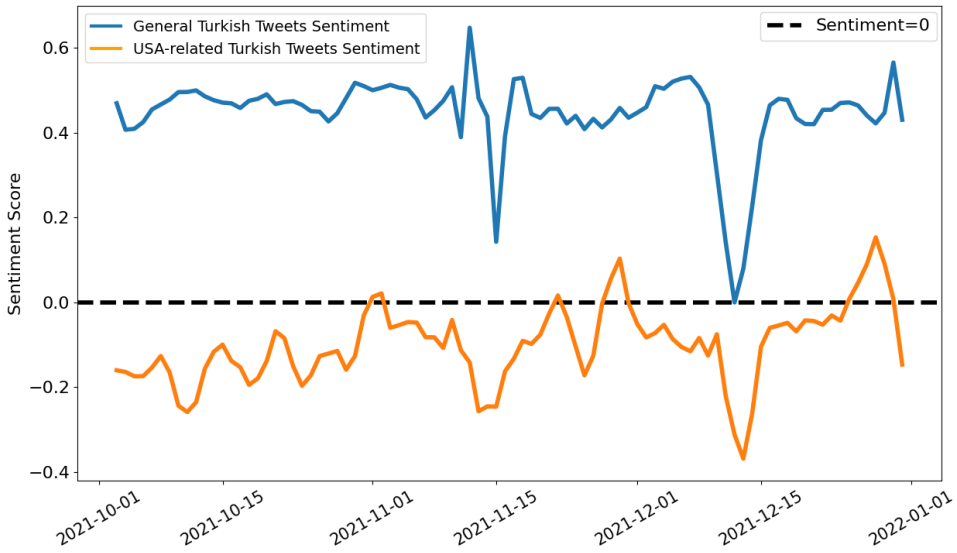


Figure 4. Daily average sentiment score of general Turkish Tweets vs. Turkish tweets related to USA.

From the emotion analysis plots in Figure 3, we can see that Turkish tweets related to the USA tend to be dominated by feelings of *disgust*, followed by *anger* and *fear*. On the other hand, general Turkish tweets are dominated by feelings of *happiness* followed by *disgust* and *sadness*, as seen in Figure 5. We can argue that generally, Turkish tweets have high feelings of happiness but tweets mentioning about the USA tend to have high feelings of disgust, confirming a negative stance against the US.

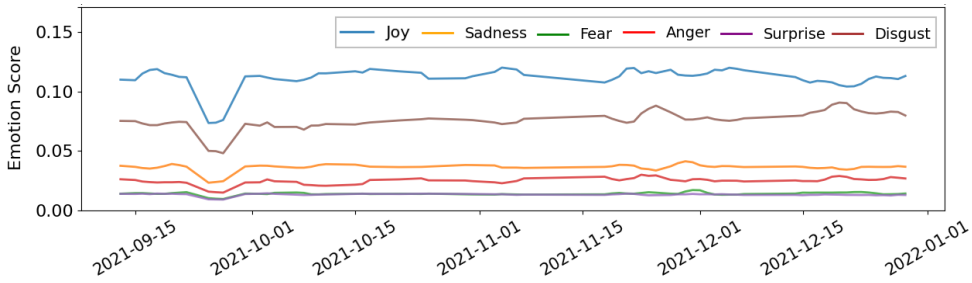


Figure 5: Emotion analysis results of general Turkish tweets

4.3 What are the common targets of these trends?

The final task is to understand why we observed such a negative stance in tweets related to the United States. In this case, we identified three targets: Trump, Biden, and the USA. The idea here is to understand whether the negative stance is against political actors or the US itself as a country. We first filtered from our original tweet dataset tweets containing these keywords. Figure 6 shows the number of tweets mentioning Trump, Biden, or US/USA daily. We then conducted TSA on these separate datasets.

Trump. The target 'Trump' is mentioned in 1,099,307 unique tweets from 2019 to January 7, 2021. Figure 7 shows the daily average probability of a tweet having negative or positive sentiment toward Trump. The bar plots also have error bars which tell us how confident we can be in our results. There are many peaks in the number of tweets about Trump and the significant events. Some of the notable events are as follows.

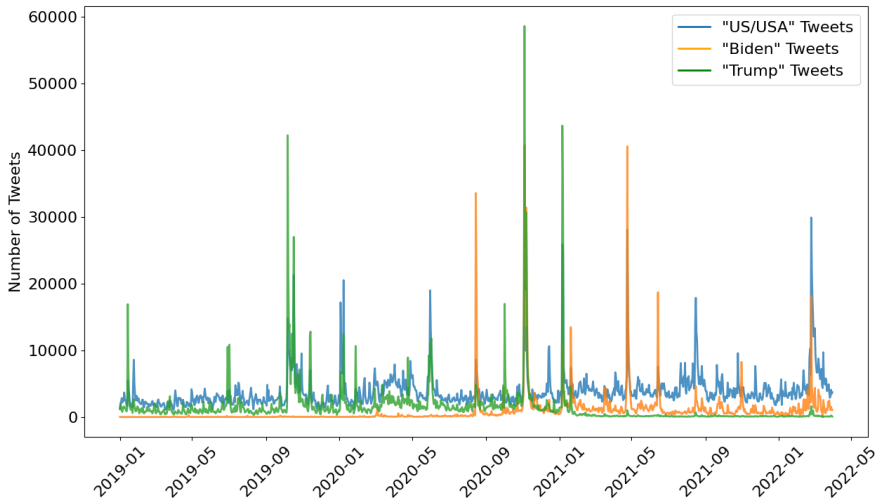


Figure 6: Number of tweets about Trump, Biden and US/USA

- January 2019: Trump vows to "devastate Turkey economically" if the Kurds are attacked (Hudson & Fahim, 2019).
- June 2019: Trump meets with Erdogan, and mentions he was open to withholding sanctions (CNN, 2019).
- October 2019: Trump announces sanctions on Turkey as a result of their invasion of northern Syria (BBC News, 2019};
- November 2019: Trump holds a press conference with Erdogan at the White House (Nakamura, DeYoung & Kim, 2019).
- Late April 2020: Trump holds a telephone conversation with Erdogan, and agrees on close cooperation on coronavirus management (Al-monitor, 2020).
- January 2021: The January 6th insurrection and response to the Capitol riots (Zaman, 2021).

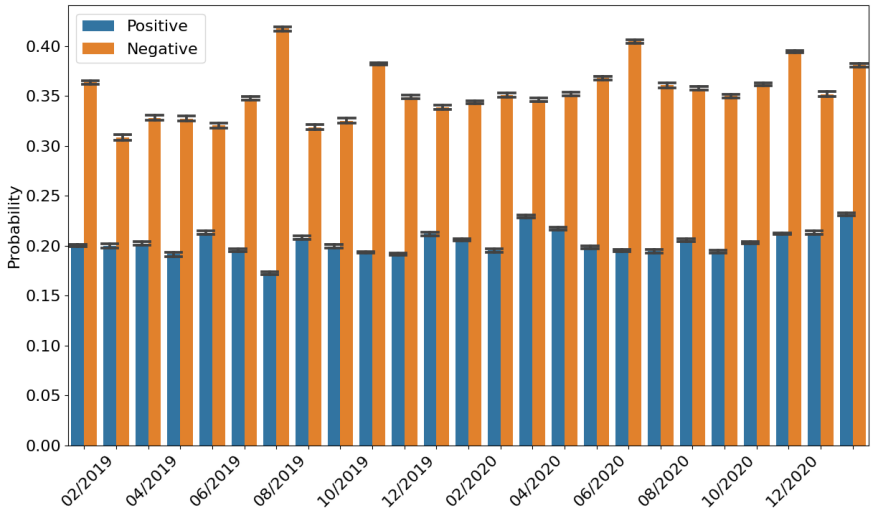


Figure 7: Probability of negative and positive sentiment toward Trump

Biden. The target 'Biden' keyword is mentioned in 1,113,909 unique tweets. Figure 8 shows the daily average probability of a tweet having negative or positive sentiment toward 'Biden'. We can see that not many tweets mention Biden in the earlier part of 2020, even though we collected Turkish tweets with the word 'Biden' in them in 2020 as well. There are many peaks in the number of tweets about Biden, and significant events associated with them are as follows.

- August 2020: Joe Biden's comments against Erdogan and the AKP in December resurface (Spicer, 2020).
- November 2020: The US election (Reuters, 2020).
- April 2021: Biden recognizes the Armenian genocide (Coskun & Evans, 2021) showing the highest negative percentage.

- June 2021: Biden at the NATO summit (Pamuk, Baczynska, Gumrukcu & Toksabay, 2021).
- October 2021: Biden meets Erdogan at G20 summit (Miller, 2021).
- February 2022: Biden came up on the day of the invasion of Ukraine, since he was issuing public statements and implementing sanctions. (Mason, Holland & Bose, 2022).

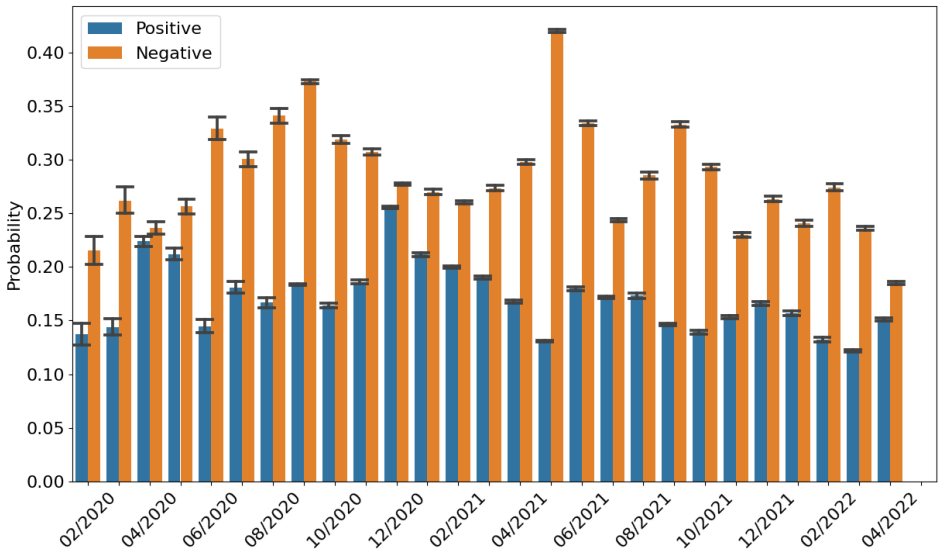


Figure 8: Probability of negative and positive sentiment towards Biden

The US and The USA. The targets 'The US' and 'The USA' are mentioned in 4,559,158 unique tweets. Figure 9 shows the daily average probability of a tweet having negative or positive sentiment toward the US/USA. Unlike the Trump and Biden targets, the US/USA target receives a much more balanced sentiment. For many of the months recorded in this dataset, there is a slightly higher presence of negative sentiment.

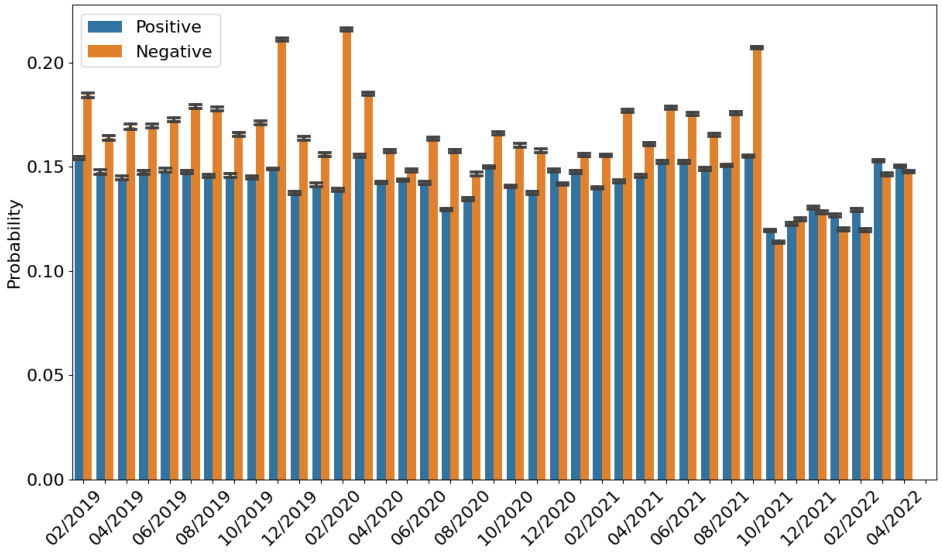


Figure 9: Probability of negative and positive sentiment towards US/USA

5. Conclusion

In this paper, we analyzed anti-American views in Turkish social media from 2019 to March 2022. We chose Twitter as our social media use case and collected data using the Twitter API. We used natural language processing algorithms like sentiment and emotion analysis to identify the general sentiment and emotions towards the USA. After that, we performed targeted sentiment analysis to understand who the common targets of these trends are. By analyzing individual targets of Trump, Biden, and the US/USA, we can conclude that Turkish Twitter users tend to hold more negative views towards political actors like US presidents. Negative sentiment against the country itself is weaker. While more research is needed, we can argue that if the US aims to reduce the Anti-American stance in other countries, a good starting point would be the narratives shared by the President.

Acknowledgements

This work is funded by the Commonwealth Cyber Initiative program: The Role of Cybersecurity in the Spread of Misinformation and Disinformation. Award# H-4Q21-015.

References

Akiş, F. A., Brakel, K., & Alimia, S. (2022, March 21). *Turkey's troll networks: Heinrich Böll stiftung: Brussels Office - European Union*. Heinrich-Böll-Stiftung. Retrieved October 9, 2022, from https://bit.ly/3ueYAGH

Alieva, I., & Carley, K. M. (2021, December). Internet Trolls against Russian Opposition: A Case Study Analysis of Twitter Disinformation Campaigns against Alexei Navalny. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 2461-2469). IEEE.)

Allyn, B., & Romo, V. (2020). Trump suspends all travel from Europe for 30 days to combat COVID-19. *National Public Radio*.

BBC. (2019, October 15). *Turkey-Syria offensive: US sanctions turkish ministries*. BBC News. Retrieved October 9, 2022, from https://www.bbc.com/news/world-middle-east-50050264

Benedictus, L. (2016). Invasion of the troll armies: from Russian Trump supporters to Turkish state stooges. *The Guardian*, *6*, 2016.

Beskow, D. M., & Carley, K. M. (2020). Characterization and comparison of Russian and Chinese disinformation campaigns. In *Disinformation, misinformation, and fake news in social media* (pp. 63-81). Springer, Cham.

Coskun, O. (2021, April 25). *Turkey says it will respond in time to 'outrageous' U.S. genocide statement*. Reuters. Retrieved October 9, 2022, from https://reut.rs/3uoV50Q

Ghahramani, M., Galle, N. J., Ratti, C., & Pilla, F. (2021). Tales of a city: Sentiment analysis of urban green space in Dublin. *Cities*, *119*, 103395.

Global indicators database. Pew Research Center's Global Attitudes Project. (2021, May 25). Retrieved October 9, 2022, from https://pewrsr.ch/3bOT2N6

Guardian News and Media. (2019, October 13). *Trump orders US troops out of northern Syria as Turkish assault continues*. The Guardian. Retrieved October 9, 2022, from https://www.theguardian.com/world/2019/oct/13/trump-us-troops-northern-syria-turkish-assault-kurds

Hamborg, F., Donnay, K., & Merlo, P. (2021, April). NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles. Association for Computational Linguistics (ACL).

Hudson, J., & Fahim, K. (2019, January 14). *Trump vows to 'devastate' Turkish economy if u.s.-backed Kurds are attacked*. The Washington Post. Retrieved October 9, 2022, from https://www.washingtonpost.com/world/national-security/trumps-vow-to-devastate-turkey-rattles-negotiations-over-syria-withdrawal/2019/01/14/1a61049c-17ff-11e9-88fe-f9f77a3bcb6c_story.html

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, *25*(3), 493-516.

Intel: Trump, Erdogan discuss coronavirus as pandemic worsens in Turkey. Al-monitor. (2020, April 20). Retrieved October 9, 2022, from https://www.al-monitor.com/originals/2020/04/intel-trump-erdogan-call-coronavirus-pandemic-turkey-covid19.html

Kirdemir, B. (2020). Exploring Turkey's disinformation ecosystem: An overview.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect?. *Political Communication*, *35*(2), 196-219.

Miller, Z. (2021, October 31). *Biden tells Erdogan that US and Turkey must avoid crises*. AP NEWS. Retrieved October 9, 2022, from https://apnews.com/article/joe-biden-g-20-summit-syria-europe-recep-tayyip-erdogan-790feb4da666e7e7ff1081dab986abc0

Momtazi, S. (2012, May). Fine-grained German sentiment analysis on social media. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1215-1220). (2012)

Mueller III, R. S. (2019). Report On The Investigation Into Russian Interference In The 2016 Presidential Election. Volumes I & II.(Redacted version of 4/18/2019).

Nakamura, D., DeYoung, K., & Kim, S. M. (2019, November 14). *Trump welcomes Turkey's Erdogan to White House, offers thanks for tentative cease-fire in northern Syria*. Washington Post. Retrieved October 9, 2022, from https://wapo.st/3nyKL2i

Narayanan, V., Howard, P. N., Kollanyi, B., & Elswah, M. (2017). Russian involvement and junk news during Brexit. *The computational propaganda project. Algorithms, automation and digital politics. https://comprop. oii. ox. ac. uk/research/working-papers/russia-and-brexit*.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Nielsen, F. (n.d.). *AFINN sentiment analysis in Python*. GitHub. Retrieved October 9, 2022, from https://github.com/fnielsen/afinn

Palmieri, L., Kaya, E., Prathap, G., Korb, A., Karahan, S., & Kavak, H. (2022, April). Investigating Disinformation Through the Lens of Mass Media: A System Design. In *2022 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 55-60). IEEE.

Humeyra Pamuk, G. B. (2021, June 14). *Biden, Erdogan upbeat about ties but disclose no breakthrough*. Reuters. Retrieved October 9, 2022, from https://reut.rs/38qj1IL

Pei, J., Sun, A., & Li, C. (2019). Targeted sentiment analysis: A data-driven categorization. *arXiv preprint arXiv:1905.03423*.

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*.

Reuters. (2020, November 8). *Turkey gives muted first response to Biden win*. Reuters. Retrieved October 10, 2022, from https://www.reuters.com/article/uk-usa-election-turkey-idUKKBN27O0EU

Schneider, K. T., & Carpenter, N. J. (2019). Sharing# MeToo on Twitter: Incidents, coping responses, and social reactions. *Equality, Diversity and Inclusion: An International Journal*.

Sonmez, F. (2019). Trump hails his own'great and unmatched wisdom'in warning to Turkey. *Washingtonpost. com*.

Spicer, J. (2020). Turkey Slams Biden's Past Call for US to Back Erdogan Opponents. *Reuters, August*, *15*.

Trump suggests he's open to withholding sanctions on Turkey. CNN. (2019, June 29). Retrieved October 9, 2022, from https://cnn.it/3nxWf6b

Welch, L. D. (2004). *Cyberspace-The Fifth Operational Domain*. Institute For Defense Analyses. ALEXANDRIA VA.

Wilks, A. (2021, August 19). *Turkey's plans for new social media restrictions threaten five years in prison for spreading fake news*. Al-Monitor: Independent, trusted coverage of the Middle East. Retrieved October 9, 2022, from https://bit.ly/3OFaamW

Zaman, A. (2021, July 22). *Turkey plans further clampdown on Independent Media*. Al-Monitor: Independent, trusted coverage of the Middle East. Retrieved October 9, 2022, from https://bit.ly/3yBttrD

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Gaps in Asset Management Systems to Integrate Railway Companies' Resilience

Jyri Rajamäki[1], Jari Savolainen[1], Rauno Pirinen[1] and Eduardo Villamor Medina[2]
[1]Laurea University of Applied Sciences, Espoo, Finland
[2]ETRA, Valencia, Spain

jyri.rajamaki@laurea.fi

Abstract: Railways and metros are safe, efficient, reliable, and environmentally friendly mass carriers. They are critical cyber-physical systems (CPS) that are attractive targets for cyber and/or physical attacks. SAFETY4RAILS project delivers methods and systems to increase the safety and resilience of track-based inter-city railway and intra-city metro transportation. Asset management plays a fundamental role in resilience management. This study analyses the gaps in asset management systems of rail infrastructure. The objective of the study is to understand the weaknesses and vulnerabilities in an asset management system that impacts resilience. The form of triangulation fashion was used for the analysis of consequences for each threat event. The research conducted included: a systematic literature review; a multiple case study review; and an analysis. The strength of asset inventory, condition inspection methods and decision-making scenarios were analysed, and as an expanded part of this analysis, mitigation actions linked to the vulnerabilities were identified. The study implies that asset management systems are most important in resilience management's response and recovery phases where the largest sudden economic implications can take place. The results of the gap analysis could be used to provide policy recommendations and standardisation efforts.

Keywords: Asset management, Cybersecurity, Resilience management, Cyber-physical systems, SAFETY4RAILS project, Rail transportation systems

1. Introduction

The EU funded SAFETY4RAILS project delivers methods and systems to increase the safety and recovery of track-based inter-city railway and intra-city metro transportation (SAFETY4RAILS, 2022). It develops a resilience-oriented framework covering the identification, protection, detection, response, and recovery of rail and metro infrastructure against cyber, physical, and combined cyber-physical threats. According to Bellini, et al. (2021), resilience is a multi-faced non-standardized concept having many different definitions and assessment methods exist, and resilience management has traditionally focused on descriptive (i.e., what happened) or diagnostic analytics (i.e., why it did happen) following an expert judgment-based approach.

Individuals and organisations have been managing assets for a long time; however, the term "asset management" started to be used more recently, since the 1980s, when private and public organisations in various sectors and industries initiated policies and procedures related to this topic. Since then, "asset management" has become a relatively new discipline, including such as knowledge, principles, scientific and practical approaches, standards and models, which have been developed across the world. It is now widely recognised that asset management is much more than an extension of maintenance practices, as the new discipline encompasses broader views than those of traditional engineering. Asset management has been gradually adopted and used by a broad range of sectors, as a systematic approach to the governance and realisation of value from the things that a group or entity is responsible for, over their whole life cycle (Zhou, et al., 2019). Therefore, it is now applied to both tangible assets (physical objects such as buildings or equipment) and intangible assets (such as human capital, intellectual property, and financial assets). Due to the different approaches in specific industries and businesses, there are various definitions and qualifying descriptors for asset management, however, these do not change the consistent core, regardless of the type and nature of the assets to be managed.

This paper comprises results obtained in the analysis of gaps in asset management systems to integrate resilience in the railway and metro context. A combination of literature and case study review is conducted with triangulation. The objective of the paper is to understand the weaknesses and vulnerabilities in an asset management system that impacts to the resilience. Such analysis supported the assessment of consequences for each threat event. The strength of asset inventory, condition inspection methods, and decision-making scenarios were analysed. As part of this analysis, mitigation actions linked to the vulnerabilities were identified.

The rest of the paper organised as follows: After this introduction, Section 2 presents the applied research methodology and research design. Section 3 deals with the literature review and discusses vulnerabilities for

improving resilience. Section 4 describes the multi-case study, the selected cases and presents its results divided into intentional and unintentional incidents. Finally, Section 5 concludes the paper.

2. Methodology

This study is addressed to understanding of weaknesses and vulnerabilities in an asset management system in a resilience management. Because the objective of the study is to understand a phenomenon, a combination of a literature review and a case study is the best option. Literature review is a summary, analysis, and evaluation of all the existing research on a well-formulated and specific question. Case studies are a way to explore a real-world phenomenon in-depth, illustrate a point, discuss the implications or meaning of an event, or compare the experiences of different individuals (Yin, 2010).

A multi-case study analysis is performed to extend existing knowledge in the research literature about gaps in asset management systems. According to Yin (2009), a case study analysis relies on multiple sources of evidence with data needing to converge in a triangulation fashion, and it benefits from the prior development of theoretical propositions to guide data collection and analysis. Here, the term "triangulation fashion" refers to the usage of multiple sources of evidence such as 1) multiple data sources; 2) among different evaluators as investigators (Patton, 1990; Miles & Huberman, 1994; Nunamaker, 2010).

The addressed form of research question during the study was "how can a vulnerability or weakness be understood in the domain of a typical infrastructure asset management system and in considering life cycle vulnerabilities under extreme events". The Unit of Analysis (UoA) was collectively discussed and considered in international meetings (n=2 work package meetings) and the most feasible and selected was as "a gap" or "a vulnerability" or as "a weakness". The paper provides the results of the gap analysis with the identification of weaknesses in asset management systems, referred to as "vulnerabilities", and improvement measures to integrate resilience, referred to the mitigation adversary tactics and techniques based on real-world observations.

The triangulation fashion is used in the analysis in the following form: 1) data sources as data triangulation (multiple literature references and case-study references); and 2) among different evaluators as investigator triangulation (n=4, researchers). The process included qualitative data from the literature and cases, which are analysed in terms of categorisation as data reduction, displays, and drawings of data (Miles & Huberman, 1994; Robson, 2002). The used form of analysis addresses to data reduction as the process of selecting, focusing, simplifying, and abstracting the used research data collection (literature and cases). The design of extended analysis is described in Figure 1.

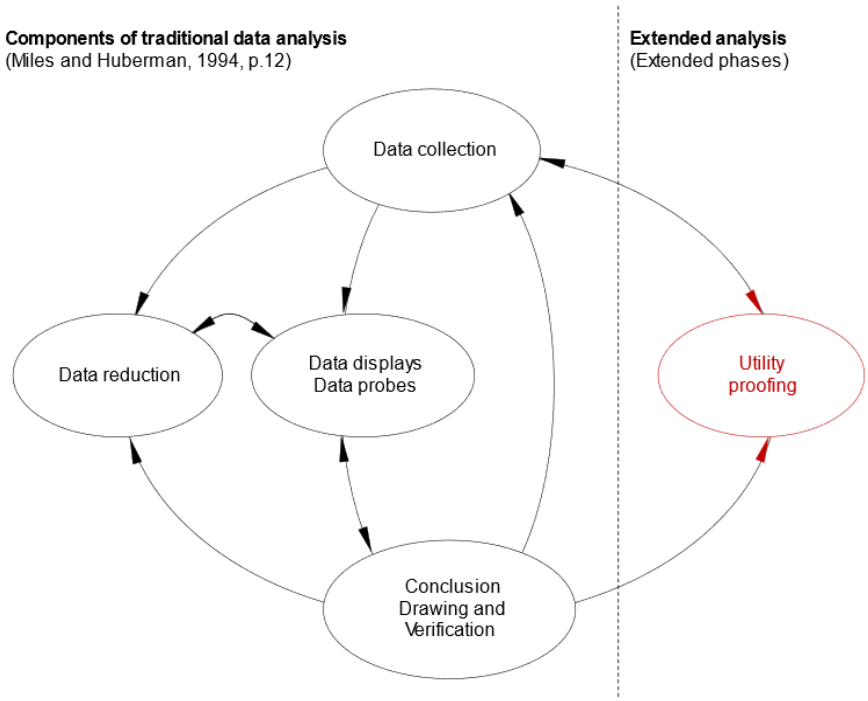


Figure 1: Research design (Miles & Huberman. 1994)

3. Literature review

The literature review was performed following the research design explained in figure 1; the Unit of Analysis (UoA) in the literature review was as “a gap” or “a vulnerability” or as “a weakness” related to the vulnerabilities and possible mitigation tactics from the perspective of asset management systems in rail infrastructure. Feasible literature was first selected as addressed reference accepted or rejected, and then selected literature was reviewed. The earlier referenced literature in SAFETY4RAILS was verified for avoiding duplicative analysis. The form of performed literature review setting was considered as including a combination of a well-conducted literature review with metadata-based analysis. The attributes of selection criteria in the literature review included: 1) metadata analysis; 2) verification of keywords; 3) existence of Unit of Analysis as a sample of evidence in the addressed literature; 4) independent three reviewers’ (n=3) collective consideration and analysis (n=5 meetings); and 5) international consortium work package meetings were arranged to ensure data quality and accuracy of relevant literature and avoiding duplicative literature (n=6 meetings). The research attributes included: research question (n=1), forming criteria (n=3); search strategy as metadata and content analysis including such functionalities as searching of title, abstract review, federated metadata search, full-text screening activities, extracting data, quality assessment, avoiding double data checking, using of searching of tactics and techniques databases for comparison, deliverable writing, and international consortium review.

3.1 Results

This section includes gathered findings of the completed literature review, including descriptions of the most frequently found literature references related to weakness or vulnerability as a sample of evidence in the literature regarding rail infrastructure asset management systems. The theoretical view of the selected analysis design is on a well-known strategic-tactical-operational-technical line, here namely functional-operative-practical-quality-human categories of attributes addressed to a typical asset management system.

From the top 40 research literature references, such as literature references (n=32) and resilience literature (n=8), the number was reduced to 6 for the third detailed extended analysis phase. Each of the 40 references was carefully reviewed taking into consideration the attributes defined in the research setting. Table 1 lists the most selected articles.

Table 1: Sample Articles

Authors	Article title	Theme studied	Sector
Emanuele Bellini, Pierfrancesco Bellini, Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Irene Paoli and Michela Paolucci	An IoE and Big Multimedia Data Approach for Urban Transport System Resilience Management in Smart Cities	Resilience management	Urban Transport
DIMECC Oy	The Finnish Cyber-Trust Program 2015–2017	Security Management	Cybersecurity
Hong, Wei-Ting, Clifton, Geoffrey, Nelson, John D	Rail transport system vulnerability analysis and policy implementation: Past progress and future directions	Resilience management	Rail infrastructures
Noritaka Matsumoto, Junya Fujita, Hiromichi Endoh, Tsutomu Yamada, Kenji Sawada and Osamu Kaneko	Asset Management Method of Industrial IoT Systems for Cyber-Security Countermeasures	Asset management	Industries
Rajamäki, J.	Resilience Management Concept for Railways and Metro Cyber-Physical Systems	Resilience management	Rail infrastructures
Junwei Wang; Raja R. Muddada; Hongfeng Wang; Jinliang Ding; Yingzi Lin; Changli Li	Toward a Resilient Holistic Supply Chain Network System: Concept, Review and Future Direction	Resilience management	Supply chain (rail infrastructure included)

3.2 Vulnerabilities for improving resilience

In this Section, the contribution of the literature research is addressed to the improvements of maturity aspects of a typical asset management system in the rail domain related to progress of resilience. While risk management considers all efforts to prevent or absorb threats before they occur, resilience is focusing more on recovery from losses after a shock has occurred as the figure below presents. This means that the main target of risk management is to protect the system from hazards outside, while the focus on resilience management is

to improve the functionality of the own system regardless of what threat it might face. In this way, resilience management covers unknown hazards that cannot be considered in risk analysis.

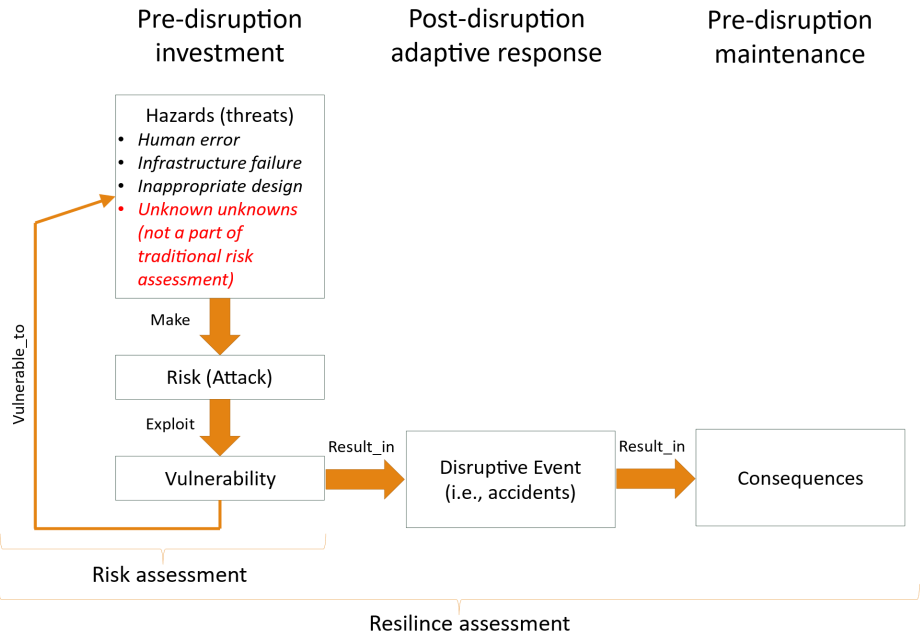


Figure 2: The relationship between hazards, risk, vulnerability, accident and consequences (adapted from Hong, Clifton & Nelson, 2022)

Previous studies usually interpret the concept of vulnerability and resilience as two sides of the same coin as shown in Figure 3. Rapidity is used to describe the time that a system requires to return to a state of normal function after a severe perturbation, such as after an intentional terrorist attack (Wang et al., 2016). The goal of resilience engineering is to improve resilience by reducing the drop in capability and speeding up recovery, and the goal of resilience management is also to learn from unwanted events and thus improve the system's capability as shown in Figure 4. Good resilience management can increase the capability of a system after a disruptive event (e.g., improved working methods after the Covid-19 pandemic).

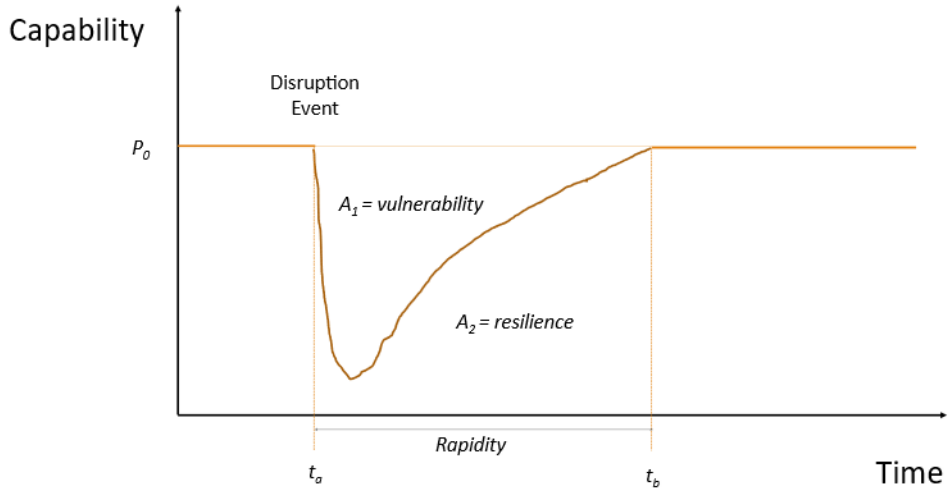


Figure 3: Concepts of vulnerability and resilience

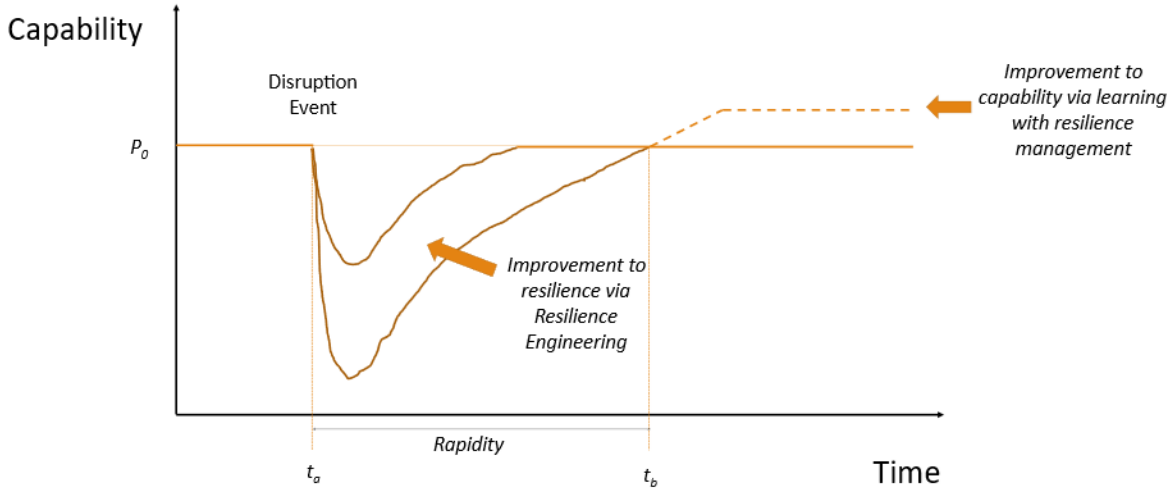


Figure 4: Resilience engineering and resilience management

The connotation of vulnerability in the context of transportation links to the reduction of capacity caused by a disruptive event, which can be estimated by any quantifiable metrics, such as total delay to passengers or the number of canceled trains. Resilience challenges of railway and metro systems are quite similar to the ones of, for example, in the healthcare sector, both being critical cyber-physical systems in which IT (information technology) and OT (operational technology) are integrated and they have a wide variety of Internet of things (IoT) sensors (Rajamäki, 2021). The appropriate asset management of IoT systems is the key to creating resilient systems. However, the timely and coherent asset management methods used for conventional IT systems are difficult to be implemented for IoT systems, because these systems are composed of various network protocols, various devices, and open technologies (Matsumoto, etc, 2021). Also, the analysis of the vulnerability of railway disruptions is seldom reviewed, and the connection between theory and policy implementation in this context is not dealt with in-depth.

The vulnerabilities identified in this literature review were checked against the ones already identified in the project. To sum up, the following vulnerabilities were found:

- Reconnaissance – Lack of monitoring of asset information and status
- Discovery of resources – Data about available resources listed in plain text
- Decision-making – Lack of quality data for decision-making (planning and control)
- Compliance Lack of compliance with asset management regulation

4. Multiple Case Study Review

The analysis of past failures is an important step toward defining the requirements of the assets management system for railway companies. The analysis will help to avoid the reproduction of known events, already experienced by railways and metro sectors, by being better prepared to face them. In this section, 94 cases were defined collecting information on recurrent, on-going and emerging threats targeting railways and metros to support the development of solutions and enable operators to enhance their resilience capabilities.

Information of the case studies was based on open sources, articles available from internet public pages. In the articles the descriptions of incidents have mainly focused on what has happened and what have been the consequences without going deeply into the reasons which have caused the incident. The press releases also do not talk about the success of the recovery phase e.g., whether the operational capability has been restored within a reasonable time. Therefore, in most of the cases the analysis of the gaps in asset management system is at most indicative.

This case study analysis was conducted in three phases:

1. Building up a data collection of the information available from internet open sources, allowing possible further studies
2. Estimation of the assets targeted in each of the cases based on expert knowledge
3. Assessment of the possible gaps in asset management system

4.1 Results

The results of the analyses are presented in Table 2 and Table 3. In the tables, the columns below the headline "Asset" describes in which asset the incidents have been targeted and in how many cases. Columns below the headline "Asset management functionality" describes in which functions the deficiencies have been assumed to exist and in how many cases.

4.1.1 Unintentional incidents

In Table 2, the unintentional incidents have been categorised based on the threat origin (natural/human) and the threat event (natural disaster, natural/environmental disaster, human failure, technical failure, environmental disaster and natural/environmental disaster).

Table 2: Unintentional incidents

UNINTENTIONAL		Asset						Asset management functionality		
		Station		Track line		Passenger rolling stock	Other	Asset status tracking	Maintenance management	Not classifiable
	Cases	Overall roof	Station facilities	Fastening system	Rails geometry	Wagon				
NATURAL	12		3		5	2	2	6	2	4
Natural disaster	2		1			1		1		1
Natural/Environmental disaster	10		2		5	1	2	5	2	3
HUMAN	11	1		4		4	2		3	7
Human failure	3					2	1			3
Technical failure	6	1		4			1	1	3	2
Environmental disaster	1					1				1
Natural/Environmental disaster	1					1				1

Natural disasters have mainly focused on solid structures like track lines and station structures. The public articles do not tell if it would have been possible to be prepared against such disasters e.g., with structural protection or if the existing protective measures have worked properly. Pre-maintenance measures for track lines might have prevented accidents in some of the cases if the problem areas could have been identified in time.

In the unintentional incidents where human has been involved the question has been of technical failure or what a person has done or left undone. For the technical failures, the gap might have been in the asset status i.e., if the asset has fulfilled the set requirements in the very beginning or if the requirements have been correctly set. The gap could have been also in maintenance if the pre-maintenance measures have not been conducted in a timely manner or in an acceptable manner.

Not classifiable cases mentioned in the table are related to such incidents which could have not been able to predict with any available asset like sudden natural landslides or human unprofessional and neglectful behaviour.

4.2 Intentional incidents

In Table 3, the intentional incidents have been categorised based on the failure type (cyber/physical/cyber-physical) and the threat event (cyber-attack, physical attack, physical attack on infrastructure, physical attack on persons and technical failure).

Table 3: Intentional incidents

INTENTIONAL		Asset													Asset management functionality		
		Station					Track line			Control centre		Passenger rolling stock					
	Cases	Electrical and lighting system	Overpass/Underpass	Station facilities	Ticketing system	Signalling	Fastening system	Overhead line	Communication system	Monito ring software	Traffic control software	Wagon	Info board	Other	Asset status tracking	IT asset management	Not classifiable
CYBER	22				15						2		2	3		22	
Cyber attack	22				15						2		2	3		22	
PHYSICAL	37		1	3			1	4	2			22		4	7		30
Physical attack	16			1				3	1			10		1	6		10
Physical attack on infrastructure	17		1	2			1	1	1			10		1			17
Physical attack on persons	2													2			2
Physical attack on infrastructure and on persons	1											1					1
Technical failure	1											1			1		
CYBER-PHYSICAL	12	2			1	1		1		1	3		1	2		10	2
Cyber attack	10	1			1	1		1		1	3		1	1		10	
Physical attack	1													1			1
Technical failure	1	1															1

In most cases, the target for cyber-attacks has been passengers' private data (Station/Ticketing system). The gap in asset management system has assumingly been in IT asset management as no evidence has been presented of failed systems functioning. Passengers' behaviour in securing their own privacy has also caused leaks in the systems as people too often tend to use the same passwords in different applications.

Physical attacks have been most common against wagons in passenger rolling stock. Means used against wagons have been obstacles on the track lines and explosives placed on the wagons in advance or carried by suicide bombers. A gap in asset management is in asset status tracking and insufficient surveillance capability (not classifiable cases in Table 3), both in technical means and in surveillance conducted by professional security personnel. Insufficient surveillance capability causes inadequate situational awareness.

Cyber-physical activities have been targeted mainly against railway/metro companies' C3 systems. As in the purely cyber-related cases studied in this analysis the challenges are in IT asset management.

In the cases that have been analysed, the cyber-physical actions have not been combined attacks, but the goal of a cyber-attack has been to make something physical to happen or vice versa the physical attack or other physical incident has caused malfunctioning in the IT systems.

4.3 Vulnerabilities identified for improving resilience

Based on the cases studied in this analysis, the biggest challenges in the current asset management system are the limited situational awareness and the lack of asset status tracking, surveillance in different information systems to reveal unauthorised intrusion attempts as well as surveillance in the real world by technical devices and physical surveillance conducted by professional security personnel to reveal physical threats. These observations are consistent with literature research, according to which situational awareness is an absolute prerequisite for resilience.

The timely and coherent asset management methods used for conventional IT systems are difficult to be implemented for rail transportation systems because these systems are composed of various network protocols, various devices, and also open IoT technologies.

5. Conclusions

The paper studies weaknesses and vulnerabilities in asset management systems that impact the dimension of cyber resilience. It supports the assessment of consequences for each threat event including a systematic literature review and a multiple case studies review. The strength of asset inventory, condition inspection methods and decision-making scenarios are analysed.

Asset management plays a fundamental role in all different phases of the resilience cycle from preparation to recovery including debriefs and audits. In the identification phase, such systems enable the identification of assets relevant to the resilience framework and in the protection phase, the management of maintenance is crucial. However, arguably, asset management systems are most important in resilience management's response and recovery phases where the largest sudden economic implications can take place. An assessment of the existing gaps in asset management systems used in these critical infrastructures will provide the necessary understanding to enforce technical and non-technical measures to integrate resilience.

In further studies with larger research material, resilience costs and time used to recover could be explored to find out how the available assets and resources have been utilised to restore the service after an accident or a criminal incident.

The paper studies existing gaps in railway asset management systems that hinder resilience integration within these organisations. Main gaps identified included vulnerabilities such as the lack of a unified system for the management both IT and OT assets, integration of the system across the various departments in the organisation, lack of revision prioritisation on key assets and manual monitoring of key assets. As part of this work, relevant mitigations actions to these vulnerabilities were also identified. The results of the gap analysis could be used to provide policy recommendations and standardisation needs. The materialisation of these gaps into a new regulatory/standardisation framework for asset management systems would provide the required mechanisms to overcome such gaps.

Acknowledgements

Acknowledgement is paid to SAFETY4RAILS Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883532. The sole responsibility for the content of this paper lies with the author. It does not necessarily reflect the opinion of the European Commission or of the full project. The European Commission is not responsible for any use that may be made of the information contained therein.

References

Bellini, E. et al., 2021. An IoE and Big Multimedia Data Approach for Urban Transport System Resilience Management in Smart Cities. Sensors, 21(435), pp. 1-34.

Hong, W.-T., Clifton, G., & Nelson, J. (2022). Rail transport system vulnerability analysis and policy implementation: Past progress and future directions. Transport Policy. https://doi.org/10.1016/j.tranpol.2022.02.004

Matsumoto, N., Fujita, J., Endoh, H., Yamada, T., Sawada, K. & Kaneko, O. (2021). Asset Management Method of Industrial IoT Systems for Cyber-Security Countermeasures. Information 12(460). https://doi.org/10.3390/info12110460

Miles, M. B. & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook. Thousand Oaks: Sage Publications.

Nunamaker, J. F. (2010). Interview with Jay F. Nunamaker, Jr: On toward a broader vision of IS research. Business and Information Systems Engineering, 5, 321–323.

Patton, M. (1990). Qualitative evaluation and research methods (2nd ed.). London: Sage Publications.

Rajamäki, J. (2021). Resilience Management Concept for Railways and Metro Cyber-Physical Systems. In T. Eze (Ed.) Proceeding of the 20th European Conference on Cyber Warfare and Security ECCWS. Reading: Academic Conferences International Limited, 337-345.

Robson, C. (2002). Real world research (2nd ed.). Oxford: Blackwell Publishing.

SAFETY4RAILS (2022). Home - SAFETY4RAILS project [online] Available from: https://safety4rails.eu/ Accessed 13.10.2022.

Wang, J. et al. (2016). "Toward a Resilient Holistic Supply Chain Network System: Concept, Review and Future Direction," in IEEE Systems Journal, vol. 10, no. 2, pp. 410-421, June 2016, doi: 10.1109/JSYST.2014.2363161

Yin, R. K. (2009). Case study research design and methods (4th ed.). Thousand Oaks: Sage Publications.

Zhou, Y. Ting, Ma, T., Zhang, H. & Chen, G. (2019). Summary - 360° of Asset Management [online] Available from: https://web.cventhen.com/event/3445a728-bd10-4019-95bd-3f39436c91f1/summary Accessed 22.11.2022.

LoRaWAN & The Helium Blockchain: A Study on Military IoT Deployment

Michael A. Reyneke, Mark G. Reith, and Barry E. Mullins

Air Force Institute of Technology, Wright-Patterson AFB, USA

Michael.Reyneke@afit.edu
Mark.Reith@afit.edu
Barry.Mullins@afit.edu

Author Note: The authors thank the reviewers for their helpful comments**.** 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: Technology is evolving at a rapid pace, and the demand for a reliable, far-reaching Internet of Things (IoT) network has never been higher. Low Power Wide Area Network (LPWAN) offers an energy-efficient, low-cost, long-range wireless communication protocol catered to IoT devices. A subset of LPWAN, Long Range Wireless Access Network (LoRaWAN), is being rapidly adopted due to its effortless integration into existing system architectures and real-world success. The Helium Network cryptocurrency blockchain's financial incentives have been used to speed up the LoRaWAN adoption and deployment in urban and rural areas by financially incentivizing gateway owners to establish a redundant network based out of their homes and businesses. In addition to ease of deployment, the Helium Network allows for enhanced security by utilizing a public blockchain ledger to verify the identities of both sender and recipient to combat packet replay and man-in-the-middle attacks. This research argues for the effectiveness of LoRaWAN and Helium Network technology fusion based on real-world examples of a robust and dependable worldwide network. Further, this research advocates for adoption and modification of this technology by the Department of Defense (DoD), to enhance environmental sensing, establish real-time tactical networks, and critical infrastructure and logistics monitoring. If the DoD chooses to integrate these two technologies with its existing IoT infrastructure; it can reliably, securely, and anonymously use LoRaWAN nodes and routers as both a long-range and backup encrypted communication network capable of supporting end-to-end encryption up to AES-128 (DoD SECRET-classification standard). The DoD could capitalize on these successes to advance information dominance in both domestic and international environments. The demonstrated performance and low adoption cost of LoRaWAN and Helium Network technologies could greatly enhance the DoD's mission of maintaining its lethality and dominance in information warfare.

Keywords: Blockchain, Cryptocurrency, Defense, Helium Network, Internet, Internet of Things (IoT), LoRaWAN, Military, Sensors, UHF Communication

1. Introduction

The Internet of Things (IoT) is an ever-evolving wave of technology that is replacing traditional sensor-related technology with low-cost, low-energy, wireless solutions designed for continual use in multiple conditions. Around the world, commercial and industrial sectors have been quick to adopt and adapt IoT technologies to expand their businesses and offer new capabilities to their clienteles. However, world governments have been slow and adverse to adopting or entertaining IoT integration for military purposes. The United States military is interested in warfighting integration with new weapon systems and communications networks but has thus far limited its investment into IoT technologies that support smart military bases, critical infrastructure monitoring, and ruggedized sensors in deployed environments.

IoT devices have distinct requirements such as radio coverage, scalability, and power consumption that can make device deployment difficult. Low Power Wide Area Network (LPWAN) (Raza, et al., 2017) is an emerging technology that is the next step in evolution of wide area, low power wireless communications. LoRaWAN, an open specification subset of LPWAN and Long Range (LoRa), operates within globally unlicensed frequency bands (see Table 1), which eliminates licensing costs and makes this technology affordable and globally deployable-- (Adelantado, et al., 2017) (Sornin, et al., 2015) similar to current long-range IoT wireless communications technologies.

When comparing common wireless communication technologies, as shown in Table 1, Figure 1, and Figure 2, LoRaWAN devices exhibit the following characteristics:

1. *Low Power:* LoRaWAN protocol's low power consumption makes it an excellent candidate for a scalable platform to support IoT devices. In fact, IoT devices with LoRaWAN capabilities consume significantly less power than those using 4G/5G or ZigBee communication protocols.

2. *Low Cost:* The low operating and manufacturing costs of LoRaWAN allow for more resources to be spent on manufacturing IoT sensors and enable the end-user to deploy more or higher quality IoT devices for the same net cost.
3. *Long-Range:* As shown in Figures 1 and 2, LoRaWAN has considerable range (3-5 km in urban areas, 20 km in suburban areas) when compared to more common wireless communications protocol such as Wi-Fi & Bluetooth. While this range is significant, it requires minimal obstructions and preferably a direct line-of-sight between access points and IoT devices utilizing the network. One unique attribute of LoRaWAN is that the signal can reflect off a limited number of buildings, water sources, and hills effectively, which further increases its range.
4. *Enhanced Propagation & Signal Penetration:* LoRaWAN operates under an unlicensed spectrum under the Sub-GHz ISM band (see Table 1). This band provides enhanced signal propagation properties that enable adequate coverage in urban or crowded areas because the signal can penetrate through natural and man-made obstructions without a great deal of signal loss.

Table 1: Comparison of common wireless communication technologies

	Bluetooth *(Bluetooth Inc., 2020)*	**LoRaWAN** *(LoRa Alliance, 2015; Libelium, 2015)*	**Wi-Fi** *(IEEE, 2020)*	**Zigbee** *(ZigBee Alliance, 2015)*
Bandwidth	1 MHz	125 kHz 250 kHz	22 MHz	300 kHz 600 kHz 2 MHz
Frequencies	2.4 GHz	433-434 MHz [EU] 863-870 MHz [EU] 902-928 MHz [NA] 915-928 MHz [AU] 2.4 GHz [Global]	2.4 GHz 5.1 GHz	868 MHz 915 MHz 2.4 GHz
Modulation	GFSK	CSS	BPSK QPSK	O-QPSK
Max Data Rate	1 Mbps	50 kbps	9.6 Gbps	250 kbps
Range (Urban)	10 m	3-5 km	100 m	10 m
Range (Rural)	100 m	20 km	200 m	100 m
Transmit Current (Max)	300 mA	135 mA	700 mA	285 mA
Transmit Power (Max)	10 dBm	14 dBm	20 dBm	20 dBm

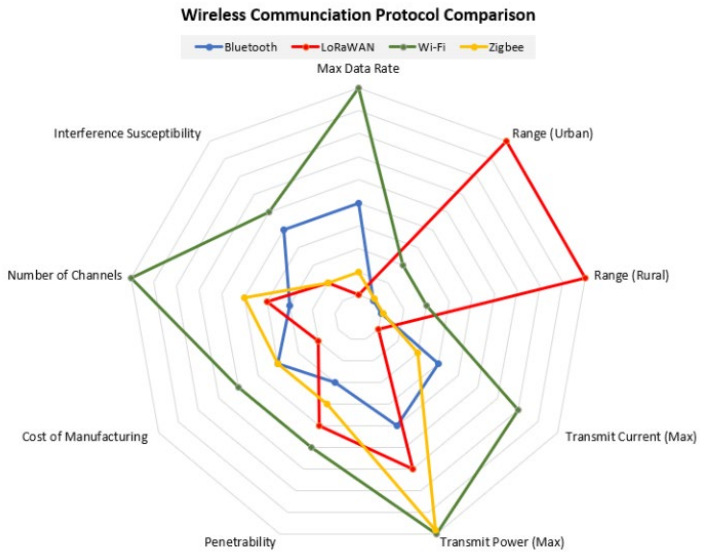


Figure 1: Radar Chart comparing Bluetooth, LoRaWAN, Wi-Fi, and Zigbee wireless communications protocol.

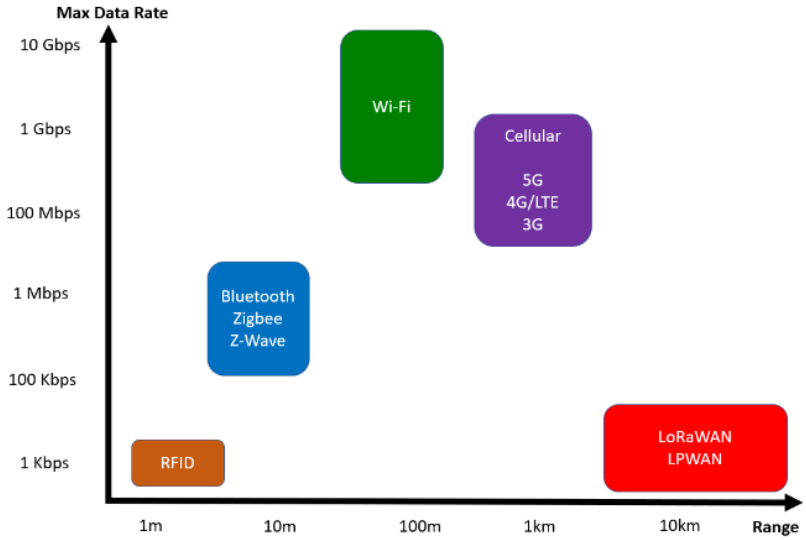


Figure 2: Max data rate and range graph comparing common wireless communication technologies

Integrating the benefits of LoRaWAN with a crypto blockchain, such as the Helium blockchain, can help people and organizations worldwide establish dependable IoT communications built around security and redundancy.

2. The Versatility of the Long-Range LoRaWAN Network

LoRa is the physical layer protocol used in LoRaWAN. LoRa features low power operation, which enables certain devices to last approximately 10 years on a single charge. Compared to other communication technologies, LoRa has a low data rate, between 20-50 kbps, depending on its modulation. However, LoRa makes up for this low-data rate by having an exceptionally long communication range (3-5 km in urban areas, 20 km in rural areas) (Adelantado, et al., 2017). The LoRaWAN network functions on a star topology with LPWAN-enabled access points, gateways, and nodes and maintains a network server connection to the Internet via IP over Cellular or Ethernet. LoRaWAN expands on LoRa capabilities by using AES-128 for end-to-end encryption and adds a frame counter to the packets (Blenn & Fernando, 2017) for verification and to prevent playback attacks. The added security and long-range capabilities make LoRaWAN favored in remote locations where routine access is limited or where sensitive data may be intercepted.

As listed in Table 2, there are three different device classes of LoRaWAN (Class A, B, & C) that can be used for different purposes in various applications:

1. *Class A:* Supports bi-directional communication between an end-device and a gateway. Uplink messages from the device can be sent at any time. The device opens two receive windows at *Rx1 Delay* and *Rx2 Delay* following the uplink transmission. The server chooses to respond in either window. Whatever window it chooses, the other one will close. Class A devices are often battery-powered and have the lowest energy consumption when compared to the other device classes. Class A devices mostly operate in sleep mode, which results in long intervals between uplinks.
2. *Class B:* Extends Class A communication by time-synchronizing receive windows for downlink messages from the server and introduces multicast capabilities. Class B devices are low latency, reachable at preconfigured times, and can receive a downlink at any time. Since these devices are periodically accessed, their battery lives are shorter, and the devices spend more time in an active state.
3. *Class C*: Extends Class B by making receive windows persistent unless the end-device is transmitting. This increases energy consumption but establishes a reliable data stream between devices and gateways. Class C devices are normally AC powered and have no downlink latency due to their always active state.

Table 2: Comparison of LoRaWAN

	LoRa Class A	*LoRa Class B*	*LoRa Class C*
Communication	Bi-directional, occasional	Bi-directional, scheduled	Bi-directional, persistent
Interval Size	Long Intervals	Long Intervals	Short Intervals
Message Type	Unicast	Unicast, Multicast	Unicast, Multicast
Power Type	Battery (high latency)	Battery, AC (low latency)	AC (no latency)
Server Relationship	Server establishes downlink, predetermined response windows	Server initiates communication at periodic intervals	End-device constantly receiving/ transmitting

As seen in Figure 3, as of July 2022, there are 173 network operators in 177 countries who provide full LoRaWAN coverage for all three LoRaWAN classes (LoRa Alliance, 2022). This coverage is more than sufficient for global DoD operations and LoRaWAN system deployment.

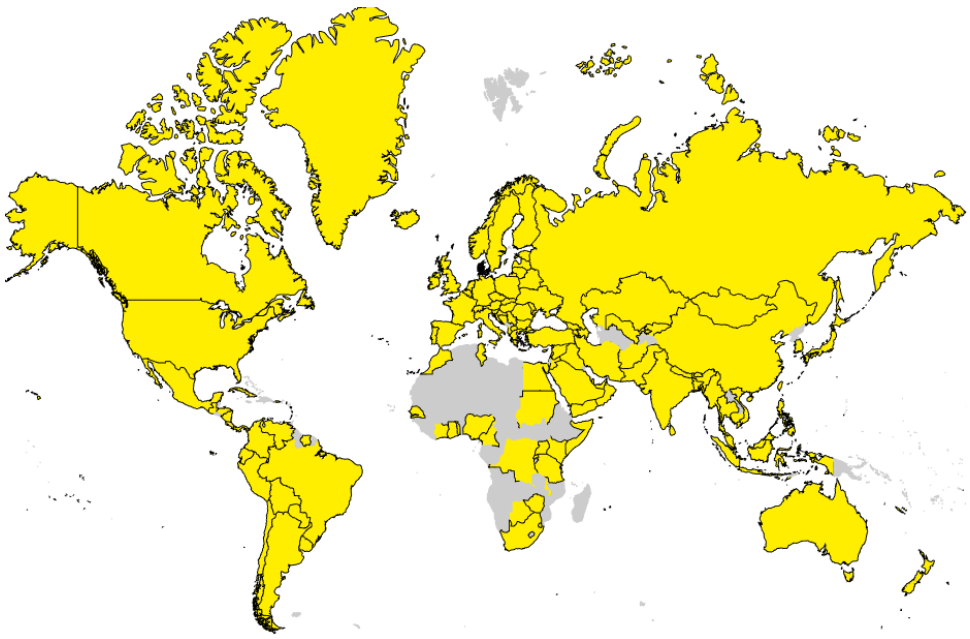


Figure 3: Worldwide LoRaWAN Coverage Map in yellow (LoRa Alliance, 2022)

The security of a communication system is equally important as its availability and reliability. For this reason, the LoRaWAN standard is designed with security in mind "to adhere to state-of-the-art principles: use of standard, well-vetted algorithms, and end-to-end security." (LoRa Alliance, 2017) The network join policy requires mutual authentication before a LoRaWAN end-device can connect to the LoRaWAN network. To prevent traffic from being altered, LoRaWAN messages are origin authenticated, integrity and replay protected, and encrypted up to AES-128 standards.

There are multiple layers of security implemented in the LoRaWAN network at both the software and hardware levels. As with most devices, the effectiveness of a system's security is dependent on the physical access of a malicious actor. If a device (e.g. node, gateway, relay) is subjected to physical intrusion, the encryption keys are stored in Secure Element, a tamper-resistant storage, which makes them extremely difficult to extract (LoRa Alliance, 2017) (Coman, et al., 2019).

3. The Helium Network, Cryptocurrency, and Blockchain

The Helium Network is a cryptocurrency-backed, wide-area networking system that uses protocol tokens for data transmissions (Haleem, et al., 2018). The Helium blockchain is a decentralized wireless network ledger designed to support LoRaWAN IoT devices. The Helium Network utilizes an open-source, standards-compliant wireless network protocol, called *WHIP* (Haleem, et al., 2018), to exchange LoRaWAN traffic for a subset of Helium cryptocurrency called a *Data Credit (DC).* DCs are spent when establishing bi-directional data transfer

between end-devices and the Internet. The process works as follows: (1) End-devices spend DCs to transfer data between the Internet, (2) Gateway operators (miners) earn tokens for providing network coverage and supporting data transfers, and (3) Miners earn Helium ($HNT) tokens for validating the coverage and data integrity of the Helium Network to prevent fraudulent proof-of-coverage or illegitimate geographical placement. $HNT tokens can then be exchanged for other cryptocurrencies or sold on crypto exchanges for fiat currencies

To establish a network proof-of-coverage, as seen in Figure 4, the Helium Network components go through the following process:

An end-device requesting to transmit data across the Internet uses a *Transmitter* to ping the Helium Blockchain. Another node, called a *Challenger*, verifies the Transmitter is a reliable node on the LoRaWAN network. Nearby nodes called *Witnesses* verify both the Challenger and Transmitter can communicate effectively. The Challenger issues a challenge to a node on the Helium Network and the Transmitter sends a beacon to witnesses. Then, the Witnesses receive the beacon and send back a proof-of-coverage receipt of Transmitter's network coverage back to the Challenger. The Challenger then validates the information and informs the Helium Blockchain of its success or failure. Lastly, if the Transmitter provided proper proof-of-coverage, they get rewarded with $HNT.

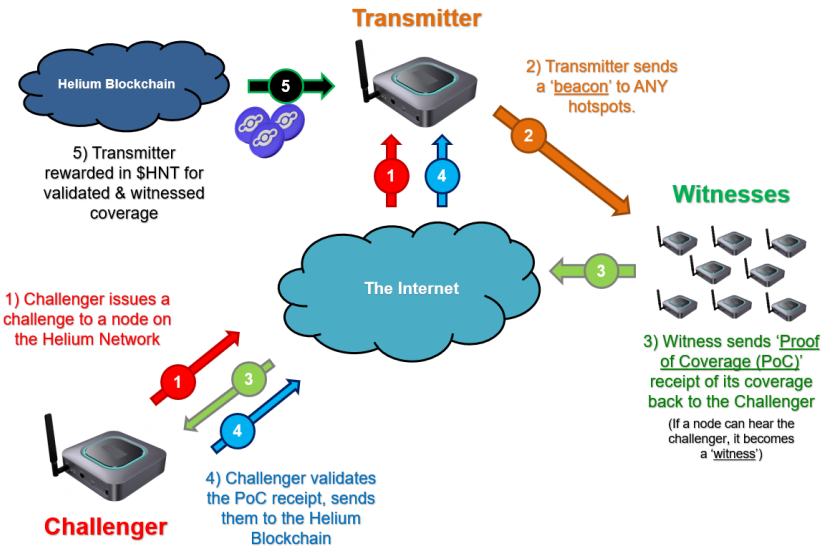


Figure 4: Helium Network Ecosystem

For an end-device to utilize the Helium Network for data transfer, as seen in Figure 5, the end-device must be able to communicate with LoRaWAN gateways and the Helium blockchain (Haleem, et al., 2018) The Helium Network uses a pay-to-communicate model, exhausting DCs, for network utilization. Without connection to the Helium Network, data cannot be exchanged between end-devices and the Internet.

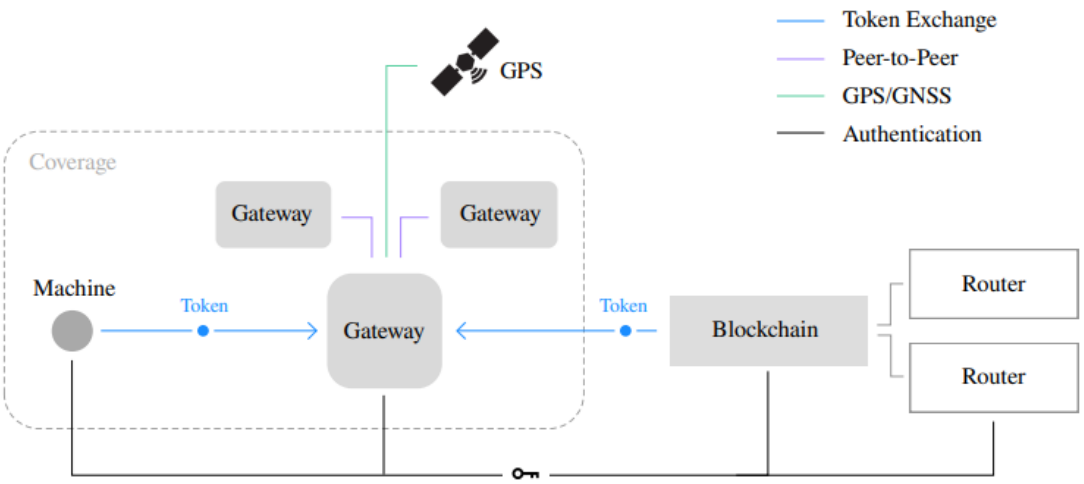


Figure 5: Helium Network Overview (LoRa Alliance, 2017)

4. Military Application of LoRaWAN & Helium Network

The military is fueled by information and demands the information be provided in a timely manner, whether in war or peacetime. By introducing an IoT ecosystem backed by or modelled after the Helium Blockchain, the DoD can create opportunities to obtain critical information otherwise deemed difficult to obtain. The three primary applications for DoD use of LoRaWAN IoT devices are as follows:

1. Environmental Sensing: There have been proven examples of accurate and timely environmental data collection such as real-time urban flood monitoring (Garcia, et al., 2015) and early earthquake detection and monitoring via QuakeSense (Boccadoro, et al., 2019). Utilizing this technology would allow for the real-time monitoring of both natural disaster precursor events and environmental parameters, which could help the DoD to coordinate timely evacuations. With this added information capability, the DoD could effectively evacuate US citizens prior to deadly events such wildfires, tornados, floods, and earthquakes that affect DoD assets. Extending the sensing capabilities in the bioenvironmental field to bolster DoD's monitoring of migration patterns, weather conditions, and environmental viability would be invaluable for global operations. To ensure DoDs smart bases and smart cities comply with environmental laws and policies, LoRaWAN allows the DoD to "pervasively monitor the air quality in urban areas by exploiting public means of transportation as mobile sensor nodes" (Addabbo, et al., 2019). Additionally, this type of monitoring can be deployed during tactical and humanitarian operations to validate information from forward-advance teams and weather personnel to enhance mission success and operational safety.
2. Real-Time Tactical Information: The long-range attributes of LoRaWAN allow IoT devices to be placed in multiple locations and on multiple assets such as personnel, vehicles, and buildings to receive, transmit, and verify data. Since there are normally no reliable sources of Ethernet in deployed or training locations, a satellite uplink would be necessary to manage the Internet uplink for these devices (Abdullahi, et al., 2019). When paired with other communication protocol combinations, such as LoRaUAV's use of Wi-Fi and LoRaWAN, real-time information such as GPS and biometric data can be successfully received and transmitted from firefighters carrying portable LoRaWAN tags (Stellin, et al., 2020). Defensively, LoRaWAN devices could be ruggedized and placed on deployed personnel to monitor human factors such as stress and vitals, as well as location to prevent friendly fire incidents (Stellin, et al., 2020). Offensively, LoRaWAN devices could be integrated into adversarial infrastructure to monitor their movement, capabilities, and identify potential opportunities for the neutralization/exploitation of their assets (Kott, et al., 2016).
3. Special consideration must be had when implementing ruggedized LoRaWAN devices in the field as they have a risk of being lost, captured, or reverse-engineered by hostile groups. A ruggedized LoRaWAN device would need obfuscated hardware and software modules to discourage and frustrate any reverse-engineering attempts. Additionally, these ruggedized devices would require additional encryption protocols to prevent signal interception and detection.

4. Critical Infrastructure: LoRaWAN's wireless signal range and broad support of multiple devices could prove to be effective in streamlining the monitoring of critical infrastructure such as electricity, natural gas, water, and sewage. (Conceição, et al., 2020) (Michaelis, et al., 2019). One such example utilizes LoRaWAN for smart infrastructure to monitor physical access to city sewer systems to increase civic response (Chaudhari, et al., 2021). LoRaWAN's ability to process Message Queuing Telemetry Transport (MQTT) messages means existing IoT infrastructure need minimal modification as they may only require a LoRaWAN antenna module, as seen in Figures 6 and 7, to connect (Fathoni, et al., 2020) (Black Hills Information Security, 2022). It is important to note that if DoD implements this technology, adoption costs and upgrade time will be extremely low.

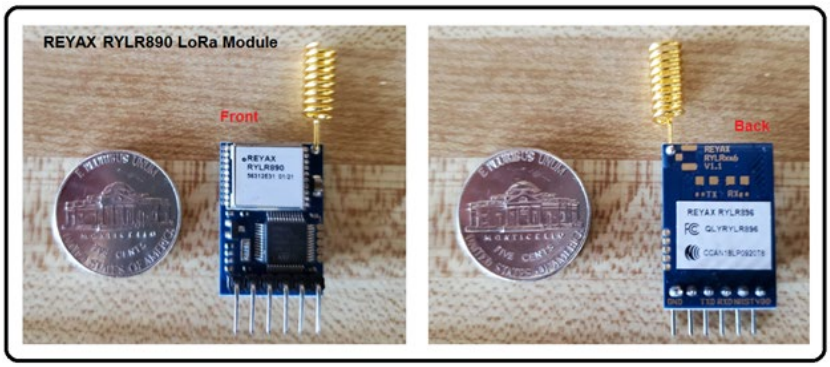


Figure 6: REYAX RYLR890 LoRa Radio Module (Black Hills Information Security, 2022)

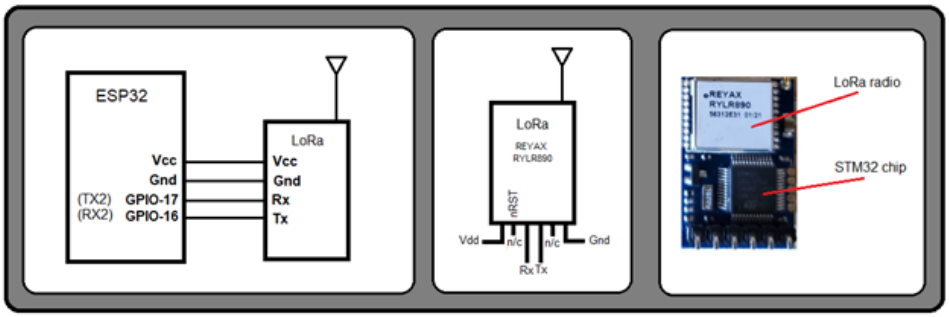


Figure 7: REYAX RYLR890 Pin-Out Diagram (Black Hills Information Security, 2022)

In November 2021, the DoD experienced a hazardous environmental disaster where over 14,000 gallons of jet fuel spilled at the Red Hill Bulk Fuel Storage Facility inside an access tunnel at Joint Base Pearl Harbor-Hickam (JBPHH), HI, USA. (The Associated Press, 2021) While the spill was being cleaned up, the jet fuel had already integrated with the Navy's water distribution system at JBPHH which serves over 93,000 people. This fuel spill contaminated residential and commercial water systems with contaminants found to be at double Hawaii's health limits for drinking water (The Associated Press, 2021). Initially undetected, the jet fuel infused water caused multiple health problems in military members, DoD contractors, and their families, and resulted in the evacuation of over 3,000 families until the problem was resolved. If LoRaWAN and Helium Network IoT sensors were installed in the service tunnels where the spill occurred and at the water distribution plant, the contaminants would have been detected sooner and the consumption of the contaminated water could have been mitigated.

5. *Logistics Monitoring:* LoRaWAN end-devices and routers could also be integrated into stationary cargo aircraft and vehicles to actively generate $HNT cryptocurrency. Additionally, end-devices could be integrated with cargo to automatically transmit logistical data to delivery and refueling locations to verify shipment integrity and provide accurate delivery timeframes. For example, an end-device could be programmed to contain the item information of a cargo pallet, be placed on the pallet or in the pallet, and be able to transmit its location, handling conditions, and package contents to an aircraft's LoRaWAN gateway or a LoRaWAN gateway at an aerial or naval port. This would automate and standardize the inventory management process while ensuring accurate delivery timeframes, proper cargo handling,

and up-to-date inventory listings. Additionally, shipping logs would automatically be kept on the blockchain.

5. Analysis of LoRaWAN and Helium Network Hybridization

By contributing to the Helium Network and adding the LoRaWAN wireless communication standard to the DoD's toolbox, we not only increase the resiliency and availability of digital communications but also introduce the risk of data exfiltration and interception.

1. *LoRa Device Classes:* With the three available LoRa devices classes, DoD would be able to selectively tailor each system to a specific environment and use-case model. Class A end-devices are excellent for environmental data collection, water monitoring, natural disaster early warning, and location tracking. DoD could embed Class A devices in training environments where access is irregular and data demands are low. Additionally, these devices could be used to detect seismic activities such as heavy vehicle use or explosions. The presence of CBRN agents could also be detected and recorded by Class A devices. A downside to using Class A devices in remote locations is the battery's integrity and ability recharge or replace an expended energy source. Class B end-devices could prove useful in smart bases and smart cities by wirelessly connecting smart metering on major utilities while having a public blockchain validate the readings and timing. This system would improve resource audit integrity and provide accurate statistics to identify system faults or utility misuse. Interconnectivity of systems remains difficult in subterranean and urban conditions, but LoRaWAN's long-range capabilities solve this problem. Class C end-devices shine where power consumption is not a concern. In addition to adding dedicated Class C devices to its IoT ecosystem, DoD also has the option could integrate a LoRaWAN module (Black Hills Information Security, 2022) into existing IoT and SCADA devices. This would introduce long-range monitoring and real-time configuration capabilities to critical infrastructure systems. This integration also provides an additional communication platform while establishing a blockchain-backed record to verify system changes, detect unauthorized access, and report system abnormalities in real-time. This persistent class of LoRaWAN devices could link geographically separated critical infrastructure nodes where traditional wireless networks cannot reach.
2. *Hybridization of Civilian and Government Gateways*: By distributing the network setup and maintenance to gateway owners, the military can save significant money and time. As discussed previously, due to the sensitivity of DoD information, enhancements may need to be made (end-point encryption and covert communications) to ensure security and functionality. While the Helium Network is maintained primarily by civilians and disinterested parties, the DoD cannot rule out the possibility of malicious actors hosting gateways to commit espionage or sabotage. The DoD would be able to assist with network reliability and increase its own data security by purchasing gateways and maintaining its own infrastructure. Additionally, DoD communications units or other government agencies could maintain gateways for cryptocurrency incentives, which could alleviate budget constraints.
3. *Pay-To-Communicate Model*: One benefit to the pay-to-communicate model is the cost-effectiveness of LoRa Class A devices since they do not transfer at regular intervals. A disadvantage of this model is that the devices cannot transmit on the network if they expend their DC balance or cannot access the Helium Network. This remains as a major hurdle to DoD integration as the appropriate amount of DCs would need to be pre-loaded to the end-devices and then replenished, as necessary. Additionally, depending on the mining capabilities of the LoRaWAN gateways, the DoD may need pay for DCs out-of-pocket to keep devices functional. The Helium Network's pay-to-communicate model requires continual maintenance and monitoring, which would add a large administrative burden to DoD personnel. With these constraints, it would be beneficial for DoD to establish a government-specific LoRaWAN network, modeled after the Helium Network, to allow only white-listed devices to connect to their network without the pay-to-communicate model.
4. *Ruggedized LoRaWAN and Sensitive Operations:* In deployed or austere environments, the DoD would need to maintain a subset of ruggedized LoRaWAN devices with enhanced security features. Ideally, the DoD would establish non-Helium Network devices to prevent the disclosure of DoD presence and communications capabilities on publicly available blockchains. Since most LoRaWAN devices operate with and without the Helium Network, we recommend utilizing this network for smart military bases/cities, critical infrastructure, humanitarian roles, and non-combat expeditionary operations. Modifications to Helium Network and LoRaWAN devices would be necessary to prior to use in combat environments to prevent inadvertently disclosing sensitive information or asset locations.

5. *Cybersecurity Concerns:* Attacks on LoRaWAN include replay, eavesdropping, packet modification, ACK spoofing, and battery exhaustion (Coman, et al., 2019). The greatest vulnerability to LoRaWAN end-devices and gateways is unauthorized physical access. "If an attacker gets access to a node, a gateway, or a server, and if strong hardware security policies are not used, the whole device or even the network must be assumed as compromised." (Coman, et al., 2019) Regarding signals intelligence (SIGINT), these attacks are less prevalent but Yang (Yang, et al., 2018) expands upon these attack vectors discussing how trivial it is for an adversary to successfully execute a message replay attack.

As seen in Figure 8, if an adversary can successfully record the wireless LoRaWAN communications between an end-device and gateway, they can retransmit a malicious version of the message that mimics the following attributes: 1) Device Address 2) Session Keys and 3) Incremented Counter Value (Yang, et al., 2018). A replay attack can be devastating as it can cause service denial and may reveal end-device configurations and locations.

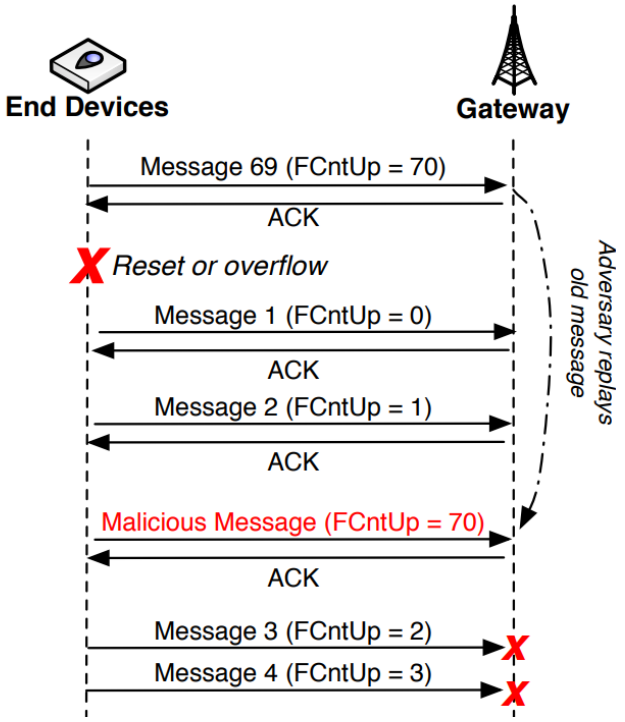


Figure 8: Example of Replay Attack for ABP (Yang, et al., 2018)

If an attacker is unable to successfully intercept and decode messages, they can use an ACK spoofing attack to selectively jam LoRaWAN end-devices. Additionally, an attacker may choose to introduce malicious end-devices into the network ecosystem with random or extreme wakeup time values causing the target's end-device batteries to be expended at a rapid rate (Coman, et al., 2019) (Yang, et al., 2018). This attack can also cause network congestion and increase the wireless transmission footprint, which would make LoRaWAN transmissions easier to triangulate and intercept.

LoRaWAN is one of the few IoT networks using standardized AES-128 cryptographic algorithms, specifically AppKey (Blenn & Fernando, 2017), for end-to-end encryption up to the DoD-SECRET level. This encryption level provides DoD the ability to transmit sensitive data if traditional communication methods become unavailable. Additionally, LoRaWAN's high physical security supports DoD's anti-tampering mission, which makes it an attractive candidate for worldwide deployment.

6. Future Work

LoRaWAN and the Helium Network are two technologies that provide secure end-to-end encryption while also building upon existing IoT communication infrastructure. This current combination limits safe use to domestic infrastructure and non-sensitive operations. In the future, the DoD may explore the possibility of creating its own blockchain for tamper-resistant communications and increased anonymity and control. Additionally, the DoD could research integrations into existing communications infrastructure for human monitoring and critical infrastructure sensing and enhancing MQTT IoT end-devices. Lastly, the DoD could also assist in building both

the Helium Network and its own proprietary blockchain-backed network to further the strengthen its information dominance and establish partnerships with both public and private entities.

To further research DoD adoption of LoRaWAN and blockchain-based technology, it would be beneficial to conduct experimentation on the reliability of communication system as well as the security of both hardware and software components. While this technology has been found favorable amongst civilian users, stressing this system for continual use in various environments is required. As with any wireless communication standard, security concerns for use with sensitive data and operations must be researched and addressed.

7. Conclusion

A unique combination of LoRaWAN and Helium Network technologies could be employed to combat IoT challenges faced by the DoD to enhance its lethality and information dominance. With a properly built IoT ecosystem, the DoD would be able to build upon existing Helium Network technologies to create a new wireless communication standard. This technological fusion provides excellent solutions for emergency management, real-time battlefield information, and critical infrastructure protection and monitoring. With minimal setup and maintenance costs, the DoD could effectively field high-quality, minimally modified devices to provide critical support to warfighters and base populi. The DoD would greatly benefit from and expand on the high potential of LoRaWAN and Helium Network technologies to create smart military bases, save additional lives, and effectively monitor the nation's critical infrastructure at home and abroad.

References

Abdullahi, U. S. et al., 2019. Exploiting IoT and LoRaWAN Technologies for Effective Livestock Monitoring in Nigeria. *Ariz Zone Journal of Engineering, Technology, and Environment,* 15(1), pp. 1-14.

Addabbo, T. et al., 2019. Smart Sensing in Mobility: A LoRaWAN Architecture for Pervasive Environmental Monitoring. *2019 IEEE 5th International Forum on Research and Technology for Society and Industry (RTSI),* pp. 1-6.

Adelantado, F. et al., 2017. Understanding the Limits of LoRaWAN. *IEEE Communications Magazine #55*, pp. 34-40.

Black Hills Information Security, 2022. *Introducting LoRa (Long Range) Wireless Technology - Part 1.* [Online] Available at: https://www.blackhillsinfosec.com/introducing-lora-long-range-wireless-technology-part-1/ [Accessed 22 July 2022].

Blenn, N. & Fernando, K., 2017. LoRaWAN in the Wild: Measurements from the Things Network. *arXiv preprint,* pp. 1-9.

Bluetooth Inc., 2020. *Specification of the Bluetooth System.* s.l.:s.n.

Boccadoro, P., Montaruli, B. & A, G. L., 2019. QuakeSense, a LoRa-Compliant Earthquake Monitoring Open System. *2019 IEEE/ACM 23rd International Symposium on Distributed Simulation & Real Time Applications (DS-RT),* 1(1), pp. 1-8.

Chaudhari, P., Tiwari, A. K., Pattewar, S. & Shelke, S. N., 2021. Smart Infrastructure Monitoring using LoRaWAN Technology. *2021 International Conference on System, Computation, Automation and Networking (ICSCAN),* pp. 1-6.

Coman, F. L., Malarski, K. M., Petersen, M. N. & Ruepp, S., 2019. Security Issues in Internet of Things: Vulnerability Analysis of LoRaWAN, Sigfox and NB-IoT. *3rd Global IoT Summit,* pp. 1-6.

Conceição, A. F., de Oliveira, L. R., de Moraes, P. & Neto, L. P., 2020. *Review of LoRaWAN Applications.* [Online] Available at: https://arxiv.org/pdf/2004.05871.pdf [Accessed 12 August 2022].

Farrell, S., 2018. Low-Power Wide Area Network (LPWAN) Overview (RFC 8376). *IEEE.*

Fathoni, H., Miao, H.-Y., Chen, C.-Y. & Yang, C.-T., 2020. A Monitoring System of Water Quality Tunghai Lake Using LoRaWAN. *2020 International Conference on Pervasive Artificial Intelligence (ICPAI),* pp. 1-3.

Garcia, F. C. C., Retamar, A. E. & Javier, J. C., 2015. *A Real Time Urban Flood Monitoring System for Metro Manila.* IEEE, s.n.

Haleem, A. et al., 2018. *Helium: A Decentralized Wireless Network.* [Online] Available at: http://whitepaper.helium.com [Accessed 21 July 2022].

IEEE, 2020. *IEEE 802.11- IEEE Standard for Information Technology.* [Online].

Kott, A., Swami, A. & West, B. J., 2016. The Internet of Battle Things. *IEEE Computer,* 49(12), pp. 70-75.

Libelium, 2015. *Libelium Waspmote LoRa 868Mhz-915Mhz SX1272 Networking Guide.* [Online] Available at: http://www.libelium.com/development/waspmote/documentation/waspmote-lora-868mhz-915mhz- sx1272-networking-guide/ [Accessed 21 August 2022].

LoRa Alliance, 2015. *Technical Overview of LoRa & LoRaWAN,* s.l.: s.n.

LoRa Alliance, 2017. White Paper on LoRaWAN Security. *LoRa Alliance,* 1(1), pp. 1-4.

LoRa Alliance, 2019. *Coverage and Operator Maps. https://lora-alliance.org/.* [Online] Available at: https://lora-alliance.org/ [Accessed 31 July 2022].

Michaelis, J. et al., 2019. Leveraging LoRaWAN to Support IoBT in Urban Environments. *IEEE 5th World Forum on Internet of Things (WF-IoT),* pp. 207-212.

Noura, H. et al., 2020. LoRaWAN Security Survey: Issues, Threats and Possible Mitigation Techniques. *IEEE Internet of Things,* Volume 12.

Petajajarvi, J. et al., 2017. Evaluation of lora lpwan technology for indoor remote health and wellbeing. *International Journal of Wireless Information Networks,* 24(2), pp. 153-165.

Raza, U., Kulkarni, P. & Sooriyabandara, M., 2017. Low Power Wide Area Networks: An Overview. *IEEE Communications Surveys Tutorials,* 1(1), pp. 1-19.

Sornin, N. et al., 2015. *LoRa Specification 1.0.* [Online] Available at: www.lora-alliance.org [Accessed 15 July 2022].

Stankovic, J. A., 2014. Research directions for the Internet of Things. *IEEE Internet of Things,* 1(1), pp. 3-9.

Stellin, M., Sabino, S. & Grilo, A., 2020. LoRaWAN Networking in Mobile Scenarios Using a WiFi Mesh of UAV Gateways. *MDPI Electronics Journal,* 9(630), pp. 1-19.

The Associated Press, 2021. *Navy Blames Hawaii Water Contamination on Jet Fuel Spill.* [Online] Available at: https://www.armytimes.com/news/your-military/2021/12/12/navy-blames-hawaii-water-contamination-on- jet-fuel-spill/ [Accessed 7 August 2022].

Tikhvinskiy, V. et al., 2017. Spectrum sharing in 800 MHz band: Experimental estimation of LoRa networks and air traffic control radars co-existence. *International Symposium on Electromagnetic Compatibility - EMC EUROPE,* pp. 1-6.

Yang, X., Karampatzakis, E., Doerr, C. & Kuipers, F., 2018. Security Vulnerabilities in LoRaWAN. *IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI),* pp. 1-13.

ZigBee Alliance, 2015. *ZigBee Specification Document 05-3474-21,* s.l.: s.n.

Blockchain Forensics: A Modern Approach to Investigating Cybercrime in the Age of Decentralisation

Saminu Salisu[1], Velitchko Filipov[2], Barry Pene [1]
[1] Bilic, London, United Kingdom
[2] TU Wien, Vienna, Austria
saminu@bilic.io
velitchko.filipov@tuwien.ac.at
barry@bilic.io

Abstract: Blockchain forensics (the use of scientific methods to manipulate data to create useful and informative descriptions of the manipulated data) takes data from the blockchain to interpret the flow of digital assets. Investigators use sophisticated data manipulation and visualization tools to identify the travel history of stolen assets. With these capabilities, chain hopping, round-tripping, and all attempts to blur transfer trails by cyber criminals become smoke screens with no effect. This paper establishes a framework for investigating financial crimes on the blockchain, starting with a brief explanation of blockchain forensics, types of financial crimes committed by criminals using the blockchain, and how they can be mitigated using OSINT's (OSINT, 2022) investigation process of data collection, data preservation, data processing, and data presentation. This paper also presents a community-based approach to financial crime investigations on the blockchain that involves the public and victims themselves contributing valuable intelligence that can be used to trace and track criminals by authorities and security agencies in the global fight against cybercrime.

Keywords: Blockchain, Cyber Crime, Digital Forensics, Data Analytics, Cryptocurrency

1. Introduction: Compliance in the Age of Decentralisation

Blockchain forensics (Cipherblade, 2022) and cryptocurrency forensics accounting involves both tracking and interpreting the flow of cryptocurrency assets on the blockchain. The data on the blockchain is publicly available, however, many wallet addresses and transactions need to be deciphered, assessed, and interpreted in each case to properly track the flow of funds and report on it accordingly.

A significant part of the global economy is immersed in dirty practices, especially in the digital financial landscape, making proceeds from illicit business transactions illegal. To the owners of these assets, the ends are all that matter. To clean these proceeds, the typical criminal would route it through several financial stops and different handlers until it gets back to its origin, but as legal and clean money. The typical money launderer loves nothing more than a process which eliminates all trails that could be traced back to the culprit. It is for this reason that cryptocurrencies and their derivatives, such as Decentralised Finance (DeFi) (Coinbase, 2022) and Non-Fungible-Tokens (NFTs) (Ethereum, 2022), have become safe havens for stolen money.

On the blockchain, criminals use a variety of methods and services that deliver money to many addresses or corporations. The assets are subsequently delivered to a target location or an exchange to be liquidated from a legal source. This procedure makes tracing laundered monies back to illegal activity extremely difficult. Now, the very novel feature of the blockchain - its anonymity and decentralised nature - constitutes a significant fear and apprehension towards the technology, especially from the traditional financial sector. Hence, the increasing development of regulations to guide and monitor cryptocurrency usage and ensure strict compliance for cryptocurrency organisations and service providers. These regulations range from registering exchanges and implementing Know Your Customer (KYC) requirements and Anti-Money Laundering (AML) (The Law Society, 2022) complaint systems by cryptocurrency organizations for the classification of cryptocurrencies.

2. Business Problem: The Cost of Blockchain Cybercrime

When cryptocurrency and the blockchain appeared on the financial market scene in 2009, it marked the birth of a technology that would threaten the stability of the traditional centralised financial system. Some of the most noteworthy features of the technology were its anonymity and decentralised design and structure, which put it outside the control of the global political elites. However, this was not a problem for the global financial body and its controlling political spine during the initial stages of the cryptocurrencies and blockchain revolution since the technologies were patronised by only a handful of weirding-looking computer geeks and were worth less than today.

"What competition could these technologies pose? What damage could they wreck?"

The world would later find out in an overly dramatic fashion what magnificent tree these crypto seeds could grow into in just a decade. In November 2021, the cryptocurrency industry reached a whopping $1.03 trillion in capitalisation, according to CoinMarketCap (CoinMarketCap, 2022). This was a decade after Laslo exchanged about 10,000 bitcoins for 2 Papa John's Pizza (Ian Wong, 2018). Today, the cryptocurrency industry has expanded to include NFTs valued at about $17 billion and DeFi valued at about $200 billion as of February 2022.

Just two short years ago, Binance (Binance, 2022) was the victim of a very swift attack that left the biggest cryptocurrency exchange at that time short of about 7000 bitcoins worth about $40.8 million at the time of the theft, making it one of the biggest cryptocurrency thefts of all times (Orme, 2019). When investigators followed the transfer trail of the Binance hack, it was discovered that about 5000 bitcoins of the total stolen funds were transferred and laundered through a service called Chip Mixer.

Another recent hack involved the $615 million dollars heist by a North Korean group on the popular play-to-earn game called Axie Infinity (Axie Infinity, 2022). The theft of Ethereum cryptocurrency from a software bridge used for the well-known Axie Infinity play-to-earn game was connected by the US Treasury Department to the North Korean hacking outfit Lazarus. The US agency went as far as placing the Ethereum addresses on the official OFAC sanction list (U.S. Department of Treasury, 2019).

Presently, billions of dollars exchange hands daily within the crypto market, and with its speed and guaranteed anonymity, the industry, by default, sends an invitation to money launderers, terrorists, hackers, drug traffickers, and other sellers of illicit materials to take advantage of the technology's ability to protect the identity of its users while allowing them to clean and funnel dirty money back into the financial system.

The global financial system needs an urgent intervention in recovering these stolen funds, creating awareness of cyber safety, and developing innovative tools to tackle this uniquely novel technology increasingly used by cyber criminals to obfuscate and launder illicit assets around the globe.

In the next section, we highlight how criminals are obfuscating funds movement on the blockchain and how they can be tracked and stopped using some of the innovative tools and services offered by Bilic (BilicA, 2022).

3. Pain Points: How Criminals are obfuscating transactions on the blockchain

3.1 Decentralised Finance

Against the design of a centralised finance system which includes third-party intermediaries and leaves a bunch of paper trails, DeFi (Sharma, 2022) eliminates the need for a middleman using peer-to-peer financial protocols.

With DeFi, anyone with a computer and internet connection can participate to lend, borrow, stake, and trade cryptocurrencies just like in traditional finance. More so, DeFi, unlike cryptocurrency exchanges, has no stringent requirements for identity verification (Moris, 2021). Thus, users' identities are protected when it comes to participating on these platforms.

The fact that DeFi platforms protect the identity of users offers an invitation to money launderers. Recent reports have revealed that about half of all the recently stolen cryptocurrencies were sent to DeFi platforms. In another report by The Block, about $192.82 billion remains locked in DeFi as of January 2022 (The Block, 2022). The argument for the rise in using DeFi for the cleaning and funnelling of illegal money is due to its present existence outside the regulatory borders (Crenshaw, 2021).

DeFi is built on blockchain technology (Consensys, 2022), hence, as a platform built on the decentralised ledger technology, transaction data can be accessed and traced. Money launderers, therefore, must employ other strategies to eliminate the digital footprints of their transactions on platforms. Some of the strategies used include:

3.2 Peel Chain

Peel chain (Hudson IntelligenceA, 2022) is used to funnel large sums of cryptocurrencies in and out of DeFi platforms to make it undetected by authorities that investigate very monstrous transactions. Peeling allows money launderers to break small transactions into copious amounts (Hudson IntelligenceB, 2022) of crypto assets and in this piecemeal, even incremental, send copious amounts of cryptocurrencies to centralised or decentralised exchanges, where they are converted to fiat currencies or other cryptocurrencies. Money launderers create awfully long and complex peel chains so that at the time the resources and processes required

to trace the transactions to their financial destination become almost impossible and tedious. For instance, an individual wanting to transfer 100 ETH undetected could perform 200 different transactions of 0.5 ETH each. This might be a tedious process, but it blends in with all other normal transactions and thus raises no red flags for the sender.

3.3 Chain Hopping

Another strategy money launderers favour is chain hopping (ChainalysisA, 2022). DeFi platforms like Change Finance, which is a multi-chain DeFi application that promotes interoperability, allow users to move their access from one blockchain network to another. Launderers often move from one chain or cryptocurrencies to another in very quick succession. This is to lose trackers and increase the complexity of their transaction history so that it becomes almost impossible to follow the transaction trail to the destination.

3.4 Mixers (Tornado Cash)

While blockchain technology makes it impossible to link individuals to wallet addresses, the decentralised public ledger keeps track of all receipts and transactions on the blockchain. Hence, transaction activities between wallets are available and can be tracked until overly broad and unrelated data are sifted, cross-referenced, and narrowed down to small variables from which to make better and informed guesses. This, to a considerable extent, is great news for crypto security and a major deterrent for money launderers on the crypto market.
Enter cryptocurrency mixers. Mixers act as digital shredders that destroy the transaction history provided by the blockchain and in the process make it impossible to track transactions effectively. An extremely popular mixer is the Tornado Cash mixer (TornadoCashA, 2022).

Tornado Cash (TornadoCashB, 2022) is an Ethereum DeFi application rumoured to have been used by hackers to launder 100 million in stolen money. With mixers like Tornado Cash, cryptocurrencies are deposited into the application by the depositing address and the deposits can only be made in specific amounts as stipulated by the application. This could be in ones, tens, or hundreds. The most plausible reason for this is that each one that contributes to the pool contributes the same amount, and when a withdrawal is made it becomes impossible to know which wallets initially deposited which and took out what.

The deposited funds are left for a brief period of time to allow the deposited funds to mix with other funds in the system so that no transaction exists as a single transaction, but a batch of transactions are layered, bunched, and mashed together to create a single compound transaction that cannot be tied to a single wallet address. The funds are then withdrawn via a different address referred to as the relayer and sent to the recipient's address, which is often different from the deposit address. This breaks the on-chain connection between the deposit address and the recipient address.

It is also important to note that most users of the platform often make use of VPN services to mask their IP addresses so that the deposit IP address is not connected to the withdrawal address. Tornado Cash mixer's ability to mash and mix transactions to remove the traces of cryptocurrencies transfers to and from different wallets makes it an especially useful tool in the hands of money launderers. In January of 2022, about 4006 ETHs were stolen from crypto.com and investigations (Reynolds, 2022) into the theft revealed that the stolen ETH had been laundered through Tornado Cash in batches of 100s to unknown wallets. According to Strom, founder of Tornado Cash, in an interview with CoinDesk, Tornado version 2 would include a cryptographic note in the transaction history of Ether during transactions to help determine the provenance of funds (CoinDesk, 2022). While this might increase the ability to track the movement of illicit funds through the platform, Tornado continues to attract huge cryptocurrency transactions because of its ability to mask transaction information.

3.5 Money Laundering using NFT

The physical art industry has always been attractive to money launderers because of its subjective price, scarcity, unpredictability, and vulnerability to manipulation. Undoubtedly, these factors, including anonymity, which is the icing on the cake, also play out in the NFT industry, resulting in the consequent embrace of the NFT industry by both legitimate and questionable business individuals. A thorough analysis of the NFT market immediately reveals its inorganic financial value explosion and the erratic movement of huge funds to purchase almost worthless digital pieces. But who cares? What is worth nothing to James might be worth gold to John, right?

Well, isn't that the right technological recipe for the manipulation and funnelling of dirty money into legally acceptable tender?

One of the popular ways money launderers utilise NFT platforms for laundering dirty money is through Wash Trading (Singh, 2022), which refers to putting up one's NFT for sale, and then purchasing it by oneself multiple times, and, in the process, manipulating the value significantly, increasing the sale price.

What money launderers do is create an account on NFT platforms, create or buy a piece of art on the platform, and wash trade" the NFT asset in their possession, using different accounts on the same platform to increase the price of the asset. When the price of the asset is significantly pumped up, they sell the asset to themselves and receive their dirty illegal money through the sale of an NFT as legitimate money which is withdrawn and circulated in society.

Chainalysis, in its report (ChainalysisB, 2022), identified 262 users who sold an NFT to a self-funded address more than 25 times, creating $8.5 million in profits. Close analysis showed that certain transactions on OpenSea showed very irregular pricing and certain accounts could be traced to have transferred funds to other wallet addresses and then used to purchase NFTs from the foundation account. These practices serve to funnel dirty and illegal money through a legitimate digital art platform back into the economy as legitimate money.

4. Solutions: Tracking Cyber Criminals on the Blockchain

The anonymity of cryptocurrencies makes it difficult but possible to uncover the identities or persons behind wallet addresses. In fact, it is this feature which makes NFTs and other cryptocurrencies attractive to bad actors.

However, a combination of blockchain investigation tools and expertise can increase precision for tracing and tracking cryptocurrency transactions and identifying their provenance. The processes of attribution and tracking money flow make it possible to reduce the numerous variables and immense amount of data involved in a single transaction to a manageable minimum so that investigators can make the most informed decisions with the information and data at their disposal.

Most people think that crypto trading keeps you anonymous. All you need to show is your wallet address or public key. An address is a string of numbers and letters, 26-35 characters long.

The truth is crypto trading is pseudonymous. This means that you cannot hide your full identity.

This brings us to the big question: Can a Crypto Transaction be traced? The short answer is YES. But it is not a 2-minute process. And it requires some expertise.

If a hack or theft occurs, an intelligence agency will need to track the hacker's wallet address. These steps are further highlighted here (BilicB, 2022).

Hash functions or strings representing individuals and wallet addresses still present a problem for de-anonymisation. One of the ways to ensure de-anonymisation (TechTarget, 2022) on the blockchain is the institution of compulsory KYC for NFT and DeFi platform users. This will make sure that wallet addresses can be linked to verifiable user information outside the blockchain, making tracing easier and more effective.

4.1 Strengthened Policy Regulations Around Using DeFi and NFTs

Such regulations would be aimed at monitoring and arresting the dangerous activities of money launderers. These policy regulations must be dynamic and flexible but with incremental strictness.

Putting in place the necessary AML system to flag suspicious transactions is very appropriate. In August 2021, hackers got hold of $610 million worth of stolen cryptocurrency. About $33 million was converted to the Tether stablecoin. The immediate discovery of the activity and swift action of Tether's chief technology officer, Ardoino, led the network to freeze the asset, making it impossible for the hacker to make away with the money. According to Ardoino, the money might have been lost forever had the Tether network wasted more time (Brown, 2021).

Recently, stolen crypto recoveries suggest that the days when cryptocurrencies provided a haven for criminals are completely over. The recovery of Poly's stolen funds, among others, suggests that blockchain investigators and forensics are improving their methods of tracking down and recovering stolen funds. Examples of tracking successes in the recovery of stolen or dirty money include the $2.3 million paid by the colonial pipeline to hackers who gained access to its computer network and Kucoin's recovery of $281 million in customer assets.

It is general knowledge that the end destination of cryptocurrencies for launderers is exchanges and DeFi platforms. Especially those with no KYC provisions. For this reason, Lisa Monaco stated that the most effective strategy for apprehending money launderers within the crypto space is to follow the money trail made available by the blockchain. The blockchain is sure to document all transactions irrespective of how complex such funds might have been routed to hide their origins and obscure their destinations.

4.2 Investigating the Blockchain

To effectively follow the money, crypto investigators use complex tools to investigate the origins of funds, the digital footprints left by the complex layers of transactions, and the destinations of funds.

Investigating sources of funds is an attempt to identify the origin of transactions, especially the address that triggers the beginning of the financial activity. As the transaction commences and goes on, it leaves footprints as it moves from chain to chain, from one exchange to the other, or from wallets to wallets.

Tracking the footprints of crypto fund flow and its analysis is often achieved using complex data mining techniques such as clustering, ownership analysis, and e-discovery:

1. Ownership analysis puts together a confluence of factors and evidence which uncovers the beneficiaries and identities behind wallet addresses under investigation.
2. Clustering algorithms, on the other hand, serve to sift, cross-reference, and identify wallet addresses that belong to the same wallet and owner. Clustering can identify thousands of addresses as belonging to a particular wallet. This helps to reduce the confusion caused by the peeling of funds and concerns the investigator with the end game in mind.
3. E-discovery helps to scrape crypto transaction and wallet data off the internet and personal devices and analytically weave the data into a giant digital roadmap which gives the investigator holistic insight. E-discovery often employs different tracing techniques to achieve its purpose.

Lastly, the destination of funds analysis is to identify the wallets that receive funds after they have been laundered through multiple layers of activities.

An important addition to crypto security, the most important one yet, which has turned the tide in favour of forensic experts, is the development and use of sophisticated software for tracking and analysis. An example is Bitquery, which produces analytical tools and software for investigations.

5. Bilic: An Approach to Blockchain AML

An all-rounded suite of tools, techniques, and visualization methods provided that empowers analysts and investigators to uncover these stolen/laundered/hacked funds would look like the Bilic framework. Our approach (BilicA, 2022) aims to help government agencies, businesses, financial institutions, and regulators detect and prevent financial crime involving crypto assets by democratising investigation through an open intelligence marketplace. Our approach adheres to the National Institute of Standards and Technology NIST (NIST, 2022).

5.1 Methodology

Data collection: The data collection phase would use sophisticated software to scrape transaction data off the blockchain.
Data Processing: The processing stage will convert the collected data into useful information used to track the origins, movements, and destinations of stolen funds.

Group Intelligence Gathering: This phase involves the combination of different intelligence departments in gathering information. This helps to cross-reference information to ensure a confluence point where the different data meets to show that information gathering is credible and would promote informed decision-making.

Reporting: The final phase involves presenting the findings of investigations for decision making.

5.2 Bilic: AML Features

5.2.1 Follow the Money (FTM) AML

FTM is our novel blockchain investigation dashboard for crypto-asset exploration, visualisations, and wallet transaction tracking. The FTM supports over 47 blockchain networks with multichain graph network visualisation support, meaning analysts can investigate Ethereum and Bitcoin wallets on the same graph.

The FTM dashboard shows the wallet transaction history including funds sent and received, In- and outbound transfers, smart contract interaction, and highlighted activities involving labelled wallets. Investigators can use the FTM tool to investigate transactions, identify nefarious wallets involved in illegal transactions, and track the funds. The FTM dashboard graphically depicts the fund's intake and outflow (see the following figure).

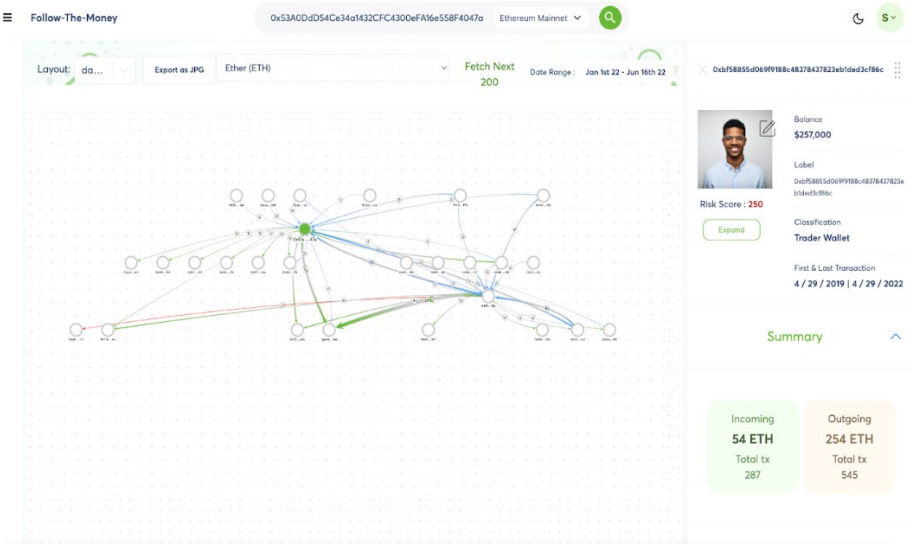


Figure 1. Showing the Follow the Money Wallet Investigation dashboard with a graph network of user transactions and wallet activities

Users can create public and private labels used to tag suspicious wallets, which can be a source of Provence for future investigators and valuable intel to both future investigators and intelligence consumers.

5.2.2 Intelligence Marketplace

The Intelligence Marketplace (NASDAQ, 2022) is an open intelligence platform that rewards data generators through funds recovery by carrying out investigations of illicit transaction data and reporting. It also offers data consumers such as intelligence agencies, and government organisations access to actionable information in a secure privacy-preserving fashion.

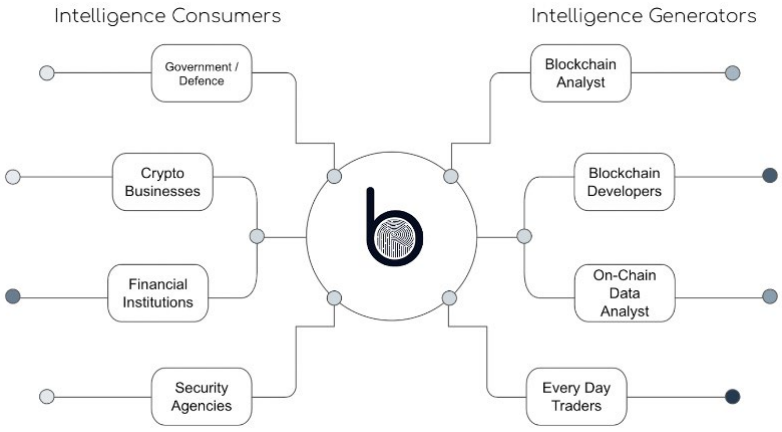


Figure 2. Showing the Follow the Money The intelligence marketplace information flow, with intelligence generators feeding and accessing data from the marketplace.

As a one-stop shop for threat intelligence information, the marketplace allows users to report security issues such as malware, hacks, scams, and fraud. Each incident report submitted through the platform is analysed, traced, and verified by our community of security professionals. After each report is examined and determined to be authentic, the information is saved on the Intelligence Data Warehouse without jeopardising the anonymity of its victims and reporters.

5.2.3 Pattern Recognition Engine (PRE)

Pattern Recognition (Rawat et al., 2022) generally utilises computer algorithms to identify data regularities and patterns in data. The PRE employs network analysis and machine learning algorithms to automatically detect patterns and irregularities in a user's on-chain activity. Such patterns and irregularities include wallets with fraudulent history, involved in fraud and money laundering, wallets linked to entities, wallets connected to the well-known OFAC sanctioned list etc.

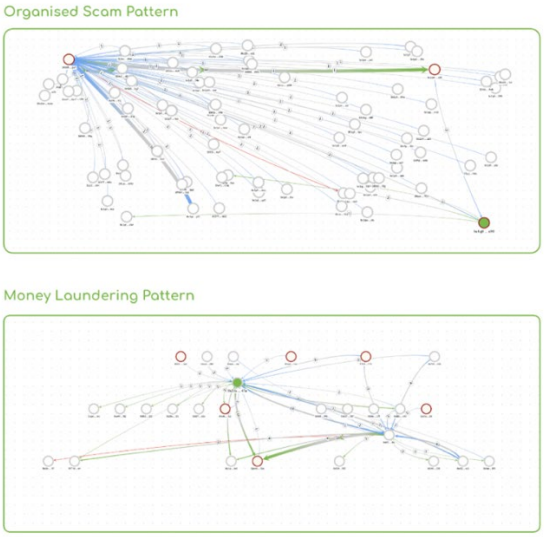


Figure 3. Showing 2 examples of pattern recognition from a wallet transaction network; the organised scam pattern above and the Money laundering Pattern below.

The engine allows the user to automatically detect a series of patterns to speed up investigations and fact-finding. Such as:

- Money laundering transactions patterns: e.g., by tracking money back to another linked wallet or carousel type of patterns. As seen in Fig. 3;
- KYC Pattern: Patterns of regular interaction with KYC platform to allow investigators to easily de-anonymise wallets;
- Mixer Pattern: Network patterns with links to mixers and tumblers;
- Organised Scam Pattern: A network pattern where funds are moved from one wallet to an operational wallet with a large number of smaller value transactions.

5.2.4 Intelligence Data Warehouse (IDW)

The IDW is our single source of truth housing large volumes of data from many sources in a centralised and consolidated data warehouse. Institutions can gain useful intelligence insights from this large data set using richness to enhance investigations, threat awareness, and decision-making. We've created a historical record over time of labelled wallets, transaction in- and outflow, off-chain media data, NFT transaction history, and wallets' association to any irregular patterns from the PRE that institutions and security analysts can use to their advantage. IDW can be thought of as the "single source of truth" for blockchain security intelligence.

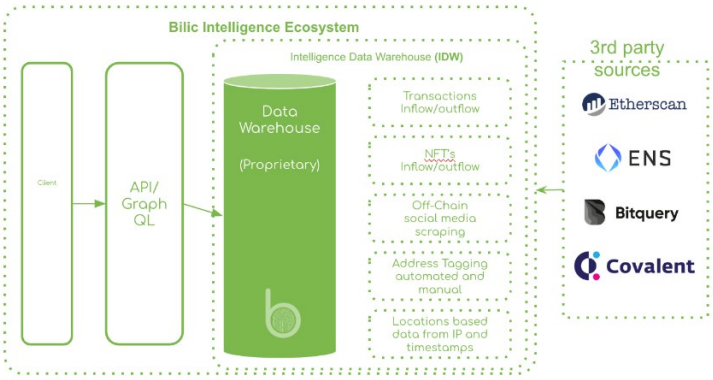


Figure 4. Showing the Intelligence Data Warehouse Architecture, with data types, storage method, providers and microprocessors clearly identified.

The IDW data is gathered from a variety of sources to create a data warehouse that no one else can match. The IDW is a critical component of the Bilic intelligence ecosystem since it allows for extensive high-level profiling, transaction segmentation, data scraping, and address tagging.

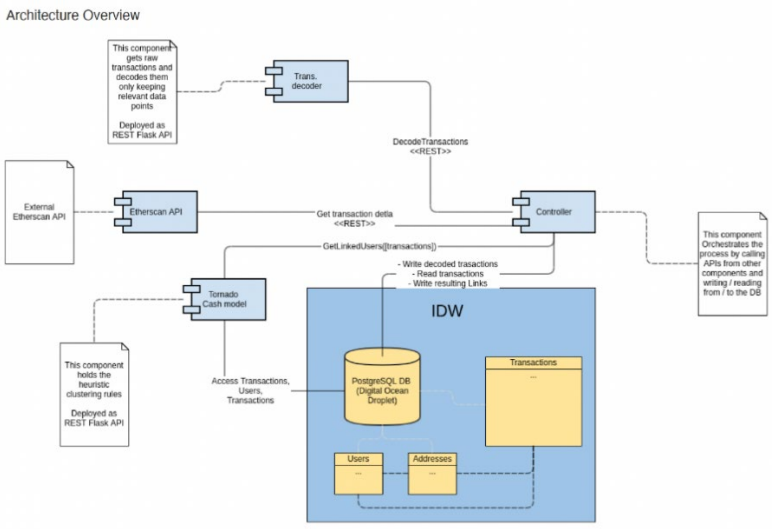


Figure 5. Showing the Intelligence Data Warehouse Architecture, with data types, storage method, providers and microprocessors clearly identified.

5.2.5 Verification

A novel AML (Notabene, 2022) tool for wallet risk factor check makes use of verification. Intelligence consumers and retail markets such as financial institutions, companies, custodians, exchanges, security agencies, trading desks, and fintech users can utilise such verification techniques as a digital shield. This solution protects consumers from transmitting their digital assets to a fraudulent wallet address in an irreversible manner. AML designs have become commonplace within the crypto industries. An example is the popular Know Your Customer exercise mandatorily carried out by some of the biggest exchanges. These AML designs enable a business to collect the personal information of users, which ties cryptocurrency wallets with real-life identities. What AML designs do is that it puts a face to wallet addresses, making it easy to identify, track and recover funds when their users commit frauds, and deter criminal activities. Users can simply query an endpoint or search for the bilic verify rating for a specific wallet and they get a detailed wallet breakdown and risk rating.

6. Conclusion

This paper has argued against the myth that cryptocurrency is untraceable and a safe haven for criminals. We have highlighted the impacts of blockchain crimes on businesses, how to find criminals attempting to eliminate

their digital footprints on the blockchain, and discussed and presented tools to track and trace these footprints along with a framework to follow when investigating blockchain-related crimes. The cryptocurrency space and its associated technologies, especially its security design, are still emerging, making it highly vulnerable to attacks. However, the days when hackers and money launderers funnel stolen money back into the economy easily are gone. There are emerging sophisticated technologies and methods for tracking and apprehending suspicious transactions on the blockchain. The increasing capability of expert blockchain investigators and crypto forensic analysts is challenged by the availability and use of transaction mixers, round-tripping, and complex layering, among other strategies used by criminals to hide their footprints on the blockchain.

References

Axie Infinity (2022). Retrieved June 29, 2022, from https://axieinfinity.com/

BilicA (2022). bilic | blockchain data forensics. Retrieved June 29, 2022, from https://www.bilic.io/

BilicB (2022). bilic | Can my transactions on the blockchain be tracked? - YouTube. Retrieved June 27, 2022, from bilic | Can my transactions on the blockchain be tracked ?

Binance (2022). Binance: Buy/Sell Bitcoin, Ether and Altcoins | Cryptocurrency .... Retrieved June 29, 2022, from https://www.binance.com/en

Brown, D. (2021). Tracking stolen crypto is a booming business - The Washington Post. Retrieved June 27, 2022, from https://www.washingtonpost.com/technology/2021/09/22/stolen-crypto/

ChainalysisA (2022). Mixing, CoinJoins, Chain Hopping, and Privacy Coins - Chainalysis. Retrieved June 27, 2022, from https://go.chainalysis.com/advanced-obfuscation-techniques-webinar.html

ChainalysisB (2022). NFT Money Laundering and Wash Trading - Chainalysis Blog. Retrieved June 27, 2022, from https://blog.chainalysis.com/reports/2022-crypto-crime-report-preview-nft-wash-trading-money-laundering/

CipherBlade (2022). Blockchain Forensics | CipherBlade. Retrieved June 27, 2022, from https://cipherblade.com/blockchain-forensics/

Coinbase (2022). What is DeFi? - Coinbase. Retrieved June 27, 2022, from https://www.coinbase.com/learn/crypto-basics/what-is-defi

CoinDesk (2022). Tornado Cash Co-Founder Says the Mixer Protocol Is Unstoppable. Retrieved June 27, 2022, from https://www.coindesk.com/tech/2022/01/25/tornado-cash-co-founder-says-the-mixer-protocol-is-unstoppable/

CoinMarketCap (2022). Cryptocurrency Prices, Charts and Market .... Retrieved June 27, 2022, from https://coinmarketcap.com/

Consensys (2022). What Is Decentralized Finance (DeFi)? | Consensys, Retrieved June 30, 2022, from https://consensys.net/blockchain-use-cases/decentralized-finance/

Crenshaw, C. (2021). Statement on DeFi Risks, Regulations, and Opportunities | US Securities and Exchange Commission. Retrieved June 30, 2022, from https://www.sec.gov/news/statement/crenshaw-defi-20211109

Dearmer, A. (2022). Business Intelligence as A Service | Integrate.io. Retrieved June 30, 2022, from https://www.integrate.io/blog/business-intelligence-as-a-service/

Ethereum (2022). Non-fungible tokens (NFT) - Ethereum.org. Retrieved June 27, 2022, from https://ethereum.org/en/nft/

Hudson IntelligenceA (2022). Peel Chain | Cryptocurrency Investigation - Hudson Intelligence. Retrieved June 27, 2022, from https://www.fraudinvestigation.net/cryptocurrency/tracing/peel-chain

Hudson IntelligenceB (2022). What is a Peel Chain? Hudson intelligence. Retrieved June 30, 2022, from https://www.fraudinvestigation.net/cryptocurrency/tracing/peel-chain

Ian Wong, J. (2018). Bitcoin Pizza Day 2018: Eight years ago, someone bought two .... Retrieved June 27, 2022, from https://qz.com/1285209/bitcoin-pizza-day-2018-eight-years-ago-someone-bought-two-pizzas-with-bitcoins-now-worth-82-million/

Moris, D. (2021). DeFi Is Like Nothing Regulators Have Seen Before. How Should They Tackle It? | CoinDesk. Retrieved June 30, 2022, from https://www.coindesk.com/policy/2021/10/19/defi-is-like-nothing-regulators-have-seen-before-how-should-they-tackle-it/

NASDAQ (2022). Marketplace Intelligence Technology Supplies time-sensitive insights and actionable information for a competitive edge | Nasdaq. Retrieved June 30, 2022, from https://www.nasdaq.com/solutions/marketplace-intelligence-technology

NIST (2022). Cybersecurity Framework - NIST. Retrieved June 27, 2022, from https://www.nist.gov/cyberframework

NotaBene (2022). What is AML and how does it apply to Crypto (anti money laundering)? | NotaBene. Retrieved June 30, 2022, from https://notabene.id/crypto-travel-rule-101/aml-crypto

Orme, J. (2019). Hackers loot £31 million in Bitcoin from Binance, wallet identified. Retrieved June 27, 2022, from https://www.techerati.com/news-hub/hackers-loot-31-million-in-bitcoin-from-binance/

OSINT (2022). OSINT Framework. Retrieved June 27, 2022, from https://osintframework.com/

Rawat, R. et al. (2022). Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence | IGI Global. Retrieved June 30, 2022, from https://www.igi-global.com/book/dark-web-pattern-recognition-crime/286781

Reynolds, S. (2022). Crypto.com's Stolen Ether Being Mixed Through Tornado Cash. Retrieved June 27, 2022, from https://www.coindesk.com/business/2022/01/18/cryptocoms-stolen-ether-being-laundered-via-tornado-cash/

Sharma, R. (2022). Decentralized Finance (DeFi) Definition | Investopia. Retrieved June 30, 2022, from https://www.investopedia.com/decentralized-finance-defi-5113835

Singh, J. (2022). What are wash trading and money laundering in NFTs? Retrieved June 27, 2022, from https://cointelegraph.com/explained/what-are-wash-trading-and-money-laundering-in-nfts

TechTarget (2022). What is de-anonymization (deanonymization)? Retrieved June 27, 2022, from https://www.techtarget.com/whatis/definition/de-anonymization-deanonymization

The Block (2022). Value Locked - Ethereum and Binance Smart Chain - The Block. Retrieved June 27, 2022, from https://www.theblock.co/data/decentralized-finance/total-value-locked-tvl

The Law Society (2022). Anti-money laundering | The Law Society. Retrieved June 27, 2022, from https://www.lawsociety.org.uk/topics/anti-money-laundering

TornadoCashA (2022). Retrieved June 29, 2022, from https://tornado.cash/

TornadoCashB (2022). Tornado cash. Non-custodial private transactions on Ethereum. Retrieved June 27, 2022, from https://github.com/tornadocash/tornado-core

U.S. Department of Treasury (2019). Treasury Sanctions North Korean State-Sponsored Malicious Cyber .... Retrieved June 27, 2022, from https://home.treasury.gov/news/press-releases/sm774

Biocybersecurity and Deterrence: Hypothetical Rwandan Considerations

Issah Abubakari Samori[1], Gbadebo Odularu[2], Lucas Potter[3] and Xavier-Lewis Palmer[1-3]
[1]minoHealth AI Labs, Accra, Ghana
[2]Old Dominion University, Norfolk, USA
[3]BiosView, Oswego, USA

issahsamori@gmail.com
godul001@odu.edu
lpott005@odu.edu
xpalm001@odu.edu

Abstract: Digitalization and sustainability are popular words within modern disciplines as practitioners each look toward the future of their respective fields. Specifically for the African continent, which is making great strides in developmental targets, those two terms are central to core aspects of policy initiatives that may foster cooperation across its varied lands and nations. One of the underlying challenges that confront Africa is a lack of strong regional integration across socioeconomic and political programs; there is value in African regions having more regional connectedness. We assess the rate of regional integration and development in Africa and discuss how to alleviate development crises that could be accelerated by deploying a sustainable cybersecurity strategy, which increasingly includes the bioeconomy and its components. This can be done through the application of Fourth Industrial Revolution (4IR) technologies such as Artificial Intelligence (AI) and modern biotechnology. This work suggests that political and socio-economic activities associated with regional integration must be seen as an all-encompassing task that transcends beyond national boundaries towards a cyber biodefense fortification and increases in 4IR technological integration. This has the aim of thereby encompassing efforts to persuade leaders to fast-track policies that seek to promote geospatial cyberinfrastructure, integrative cybersecurity considerations, cross-border digitalization programs, and increased need for cybersecurity research and education, with mindfulness towards education and further integration of mindful automation. In conclusion, a model of integrative security is proposed for Africa.

Keywords: Africa, Cybersecurity, Machine Learning, Cyberbiosecurity, Biocybersecurity, Public Health

1. Introduction

Cybersecurity policies or notional counterparts are commonplace in both the Global North (GN) and Global South (GS). Sub-policies hoping to deter cybercrime may not be found explicitly in national cybersecurity plans like that of Rwanda. However, there is a need for clearer and more comprehensive policy language that can allow deeper discussion into the contexts in which they may be enacted or pursued. Some policies, extending from emergent trends in technology development cannot emerge easily, especially when they are of hybrid fields that depend on infrastructure that is believed to belong to different domains. Herein exists the problem of biocybersecurity (BCS) and cyberbiosecurity (CBS). These describe the bridge between biosecurity, cybersecurity, and cyber-physical security and are especially relevant within the bioeconomy wherein ample cyber-physical interfaces in critical infrastructure exist; Vulnerabilities here carry national security implications as the tools to exploit them are increasingly easier to access (Murch et al. 2018; George 2019). Threats that can occur at this intersection concern the weaponization of linked biological and digital information and products that can critically derail bio-economies and the populations that depend on them (Duncan et al, 2019; Richardson et al.,2019). Attacks from their weaponization can stall development in GS countries and prevent the bridging of development gaps. To anticipate and counter BCS/CBS threats require high levels of specialization in biosecurity and cybersecurity. As biological and cyber systems further intertwine, the attacks by malicious actors at this intersection could appear and grow, accompanying the growing wave of GS cybercrime. Further, attacks through methods such as epigenetic logic bombs, medical device-driven person DDOS-ing, intra-implant surveillance, and multiplication of attack vectors via drones, mosquitoes, or other means, may allow for novel horrific means of civil disruption (Millett and dos Santos, 2019; Turner, 2019; Vinatzer et al., 2019: George, 2019; Richardson et al.,2019; Potter, 2020). The types of attacks, from the biological end, can be expanded by the complexity of the biology that is exploited since aspects of biology can serve as unique interlocks in system vulnerabilities.

What this means for nations that possess high biodiversity, such as those in Africa, is that they sit uniquely in terms of potentially weaponizable aspects of biology through consideration of pathogen storehouses in diverse environments (Samori et al, 2023). This prompts us to consider a new age of MAD in the context of Synbio and Cyber, provided that threat actors become more competent with both (Samori et al, 2023; Corlett et al., 2020;

Edwards and Revill, 2016; Gronvall, 2019). The means of deterring these types of exploits would be beneficial to develop nations within the African Union; therefore, a policy of BCS/CBS deterrence is offered for consideration. Ideally, one could suggest alternatives to BCS/CBS deterrence, but no alternatives exist. Academically, the concept of BCS, CBS, digital biosecurity, and or any variant has only existed with a degree of clustered rigor in the past decade and attempts to address BCS/CBS is quite new (Peccoud et al 2018; Murch and DiEuliis 2019). It is common to hear military, intelligence, and even politicians speak of cybersecurity, but "cyberbiosecurity" is yet to be treated with equal importance even in the Global North. This potentially highlights an issue of how much of a head start the Global North has in targeting unique GS holdings with respect to the value of their flora and fauna. A deep question exists in how holistic a country can examine itself, but that requires unity and a strong command of resources. In this work, the AU is used as an example of a partial GS commentary. Further, it is focused on the African continent with an emphasis on the continent's unique challenges regarding the Central African Nation of Rwanda as a partial study. Based on this background, the article discusses BCS/CBS and specific BCS/CBS deterrence policy for Africa.

2. The Greater AU and On Rwanda's Experience in Cyber-Deterrence

The creation and growth of the internet have brought forth great wealth and development to all parts of the world, albeit unevenly. As such, each economic sector, in each country, has used it to improve marketing and distribution. Governments have used it to both implement policy and monitor, to bring security to their nations. This has led to the increasing popularity and necessity for securing the use of the internet and protecting those who use it. At times the control of the internet has been used to prevent insurrections, stop improper exchanges of data or resources, and monitor malicious actors, especially nation-states as is evident in recent times has occurred, but a careful in-between has been preferred leading to the discussion of alternate means (Gohdes, 2020; Marchant and Stremlau, 2020). This is doable, as the options are many but necessary to ponder, as Cybersecurity is a vast domain, with implications affecting both national and international security.

Despite boasting more than 1.2 billion individuals, the 54 nations of the continent of Africa have by large been minimized in conversations surrounding cybersecurity much of this can be ascribed to comparative or uneven gaps in economic or military development, a shortage in professionals, low-penetrance, and more, in addition to historical prejudices and colonial hegemonies (Adomako et al 2018). Even with strides in development and innovations fostered among emigrants and individuals within that have initiated revolutions in technology abroad, it is still common for individuals in developed nations to ignore the evident progress, but instead, let their biases rule their judgment (Gewald et al 2012; Endong and Obi, 2020; Olubela 2018; Williams, 2020). Regardless of these biases, these young African nations are pressing on and aiming to catch up to many of these older and developed nations despite the hurdles that exist. This resolve is necessary as much cyber infrastructure remains underdeveloped and cybercrime persists in ways that can undermine cyberspace autonomy in many areas of the continent, which poses risks for development (Abrahams, L., & Mbanaso 2017). These concerns can be countered through strengthening intra-continental policies in cybersecurity, such as deterrence, as their growth is without doubt linked to their ability to cement their grasp on this important and explosively growing field. Especially in the case of fast-developing countries like Rwanda, the protection of government-administered services and private stakeholders is paramount (Nkusi, 2017). As of 2016, just about half of the continent's countries have enacted laws against cybercrime, penetration exists at slightly less than a third of the continent's population, and cybercrimes cost the continent over a billion, annually (Adomako et al 2018). This will need to change as Rwanda and other African countries continue to develop. Rwanda was chosen for this paper. It is an example of a young African country that has experienced not only rapid growth and development but while landlocked and emerging from a relatively recent civil war. Their progress has lessons for other GS nations to study, imperfections considered. This paper briefly examines part of its National Cyber Deterrence Policy, based on its 2015 National Cybersecurity Policy (Republic of Rwanda, 2015).

The policy in question is that of deterrence, within the scope of cybersecurity. The policy was developed by Rwanda, to outline and establish a national cybersecurity policy. It is part of a trend by other developing nations to catch up to development in policy with that of others (Adomako et al 2018, Lewis, 2014; Nkusi, 2017; Republic of Rwanda, 2015). In terms of how the policy is applied, insufficient information is given, owing to scant literature in the core document, the field, and from the government; the document itself referred to broad means of establishing capacity (Makanda et al 2017; Republic of Rwanda, 2015). However, there is a spectrum of potential ideas, which could prove useful. For example, from the positive and relatively inexpensive end, nations could make use of "cumulative deterrence" (Tor, 2017). Within, nations diplomatically pool resources to monitor and discourage acts. This is attractive as it allows funds normally appropriated for solo deterrence to be diverted

elsewhere in areas that possess a more dire need. However, this may require concessions within the nation of Burundi, with whom it split, that could cost Kagame popularity and thus internal power; careful steps are needed to avoid inflaming tensions that still exist (Ndayisaba, 2020). Further, a policy of cumulative deterrence might also require further learning on its other adjacent neighbors, but this might collectively achieve the same burden, but this remains to be seen as Rwanda develops its deterrence capacity. Crosston (2011) noted that many experts believed that even the "advanced" modern powers were unprepared for sophisticated, nation-state sponsored cyber-attacks and suggested that perhaps a mutually assured debilitation would be the best hope. For much of the African continent, this remains challenging in a conventional sense since considerably more of its internet penetration has been linked to the use of mobile phones versus that of fibre cables or the like (Makanda et al 2017). For example, in a study on Malawi, utilizing data from Malawi's 2015-2020 Strategic Plan, Makanda et al (2017) found that it is common for phone use across the content far outstrips the fixed broadband lines that exist. While this has allowed for mobile banking revolutions that have spread worldwide like that of the M-PESA system in Kenya and similar, also highlights an exploitable vulnerability and option in offensive capacity for Rwanda to pursue (Gewald et al 2012; Mas and Radcliffe 2010). This is an option that would require building an offensive capacity to dent mobile networks across the continent, which could backfire as no weapon remains unnoticed for long and could trigger an arms race in an offensive mobile capacity that could spread internationally (Dunham, 2008; Lakhdhar et al 2017). This could also damage Rwanda's image, internationally, reducing its ability to obtain assistance, utilize its soft power, and eventually threaten its pace of development, as with the case of North Korea, which has received scorn for not just its nuclear power development but also its capacity in offensive cyber operations (Boo, 2017; Haggard and Lindsay, 2015). However, some offensive capacity could prove helpful in countering the growing capacities of other nations. A dynamic balance could be in its best interest, for which some scholars like Iasiello (2014) may concur. His work noted the transience of many cyber-deterrence policies and how the field of cybersecurity is not static and requires adaptable and dynamic plans (Iasiello, 2014). Rwanda would be wise to tread carefully in the implementation of the deterrence policy.

In terms of how the policy fits into national/international policy/strategy, one can answer that it does well. According to its 2015 policy, it is in the process of building and strengthening its capacity, which may show remarkable progress by the release of its next edition (Republic of Rwanda, 2015). While scepticism in meaningful policy implementation may exist due to a shortage of professionals as noted by Adomako et al (2018), there is a long-running optimism as noted by Goodman (2010) who believes that deterrence is easier in practice than is theorized when grounded in the physical realities that shape any idealized plan. However, to make good on such optimism, there will need to be an upsurge in expertise to match cybersecurity, broadly, and internally. These concerns have future and ecological implications as well. Mueller (2020) notes in Cyberbiosecurity, expertise is difficult to come by which may pose difficulties on the biological end. This especially deserves mention as Central Africa boasts high biodiversity as Fa et al (2014) note, and countries like Rwanda may find value in leaning on sustaining that aspect when considering dividends that it may play in its bioeconomy. Numerous authors such as Peccoud et al (2018), George (2019), and Turner (2019) have commented on how the bioeconomy is linked to cybersecurity through cyber-physical workflows present in bioprocessing and how it inextricably poses national security, and thus international security implications.

3. An Overview of Regional Actors and Processes

In terms of the policy overview, and how it serves to resolve the existing problem, the issue is both simple and complex, in that no broad policy exists. Through proposing a policy, we address an emerging problem whose worst threats to human safety are in the future. With regards to BCS/CBS, interdisciplinary experts with some technical knowledge have been working to propose strategies that take best practices from biosecurity, cybersecurity, human factors, and related, but these proposals ultimately yet to find their way to the ear of governing authorities and much-less lawmakers. This context finds difficulty in that a sufficiently educated lay population, legal base, and governing base are needed for effective communication, the passage of law concerning, and the implementation of helpful policy. This problem is exacerbated by the fact that a comprehensive cybersecurity policy agreement itself has not been enacted in cybersecurity for the AU (Turianskyi, 2020). Further, the AU, much less than the EU, or the Global North is ready for a comprehensive BCS/CBS deterrence policy. At the same time, experts such as George (2019) have already pointed out and indicated that BCS/CBS possesses national implications, which also leads to international implications. The Global South is targeted for such a proposal instead of the Global North in that countries in each are interlinked through trade and history. What is allowed in the less prosperous areas, can eventually filter upward, much like increasing cybercrime that is spreading throughout GS, in part due to weak policy and enforcement, coupled

with economic situations that push individuals within more easily to crime (Baylon and Antwi-Boasiako, 2016). That said, consideration of the least equipped party is also important in the international scope as it eventually engages the most equipped parties. Everyone is eventually affected and is an actor. Next exists the question of how to bind actors. The AU is far from finishing the process of adopting a continent-wide cybersecurity agreement, the Convention on Cyber Security and Personal Data Protection (Turianskyi, 2020). It is difficult to imagine that they would be able to string together a BCS/CBS agreement at this point, with that in mind. Upon the day that they can consider their unique eco-biology, they should consider how each country's diverse flora and fauna can be exploited and protected in light of BCS/CBS. Within, they possess deep strategic resources that could be better protected through similar legislation. Such could also bolster conservation efforts across the continent that have been battered due to poaching and sacrifices of biodiversity for their economies.

4. Potential Impact of AI on Biocybersecurity

Artificial intelligence (AI) has seen remarkable progress over the past decade. This could be attributed to the rise in data generated and the amount of computing power available (Xu et al., 2021). Generative AI models, like Generative Pre-trained Transformers (GPT) (Brown et al., 2021) and DALL-E (Ramesh et al., 2022), are able to generate convincing images and text that are indistinguishable from what humans could generate. Some large language models (LLMs) are able to generate coherent convincing texts on subjects that seem like they were written by experts. This becomes worrisome when dealing with biological data. As more biological data (be it a newly discovered toxin or the latest variant of an infectious pathogen) is discovered, these generative models can get a better understanding of the makeup of these toxins and infectious pathogens and the underlying mechanisms based on which they operate. A system with such an understanding of biology can be used to generate lethal bioweapons that could have little to no weaknesses. This is alarming given that Africa is home to diverse biological species, which is good for an AI system when learning. AI can assist in accelerating crimes against biosecurity and cyberbiosecurity.

5. Implementation

Implementation ideally should be carried out by all African countries, and this would be strengthened with the increase in signatories by each additional country. Implementation would require stronger cybersecurity practices across the continent, in pace with its rapid connectivity, but also stronger biosecurity investments with appropriate mirror investments in infrastructure and practices. An easy benefit is that it can strengthen the growing efforts in combating epidemics from country to country, and more efficiently help contain epidemics.

Massive investments, unity, and anti-corruption measures would be required. Not all countries of the Global South are stable and possess the financial capacity to undertake endeavours to strengthen their cyberbiosecurity. As a result, incremental steps such as regional partnerships and foreign aid might be a way forward as each country finds its path. Each country will need to consider its unique, vulnerable resources to protect. That will require a means of increased scouting, public learning, and accounting beyond mere conservatorships of wildlife preserves and zoos.

This endeavour would be costly for any country considering the costs of current biotechnology. DIY efforts have shown that some costs in scouting can be reduced and distributed among citizens; this especially goes for tasks such as scouting biological resources and education, which are among the heaviest work (Bello et al., 2020; Sarpong et al., 2020; Kosaki, 2020; Wamuchiru and Moulaert, 2018; Kong and Bakker, 2018; Gruber, 2019). Even for advanced economies, there are not enough specialists to go around. Consequently, GS policy successes could entail benefits for the Global North. Community Bio groups have shown that biotechnology access does not need to be hyper-specialized, and education costs can be lowered as well, increasing access (Bello et al, 2020; Tylecote, 2019).

Communities in the Global South can identify, and catalogue unique aspects of their country's life forms, especially guided by local experts. Al can be adapted to sort through and categorize data. While intensive, this is important as such work affects everyone allowing for the improved mapping of both beneficial resources and pathogens among both humans and their environments. This has been addressed by Samori et al, 2023. With more biological resources mapped and improved education surrounding it, from the bottom-up, comprehensive biosecurity and thus BCS/CBS policy can more easily enter the mind of politicians to craft as citizens press politically from their increased appreciation and stake in their resources (Castro and New; 2016; Falzon et al, 2020).

6. Limitation and Conclusion

To summarize, Rwanda, like many other developing nations, has begun taking cybersecurity seriously and is developing both its infrastructure and means of bringing its cybersecurity policy up with international and intercontinental standards. In terms of deterrence, it has many directions in which it can go, but the direction in which it treads must be done at a speed that allows it to stem cybercrime and provocations by neighbouring and faraway nation-states. Further, its policy must carefully balance neighbourly burdens without cost to internal stability which could arise due to too many concessions or over-investment. In the absence of sufficient detail from their 2015 policy, international voices were added to this work in order to aid the discussion of the broader framework of cyber-deterrence concerning the case of Rwanda.

Adding BCS/CBS considerations to a national deterrence policy or Defend Forward initiative seems to be the most practiced deterrence policy at this time. (Deeks, 2020; Palmer et al, 2021). As such, a country can divert military funds to its initiative and gradually sharpen such a policy to expand with allies (Smeets; 2020). In general, BCS/CBS deterrence policies will need to be developed in the future as biology increasingly finds its way into conversations and arsenals of cybersecurity. For security, countries need to have adequate BCS/CBS infrastructure and policy in place before the biological equivalent of Metasploit becomes commonplace. Pursuing this policy appears preferable and may be the only one that reasonably exists.

References

Abrahams, L., & Mbanaso, U. (2017). State of Internet security and policy in Africa. The International Institute of Tropical Agriculture (IITA) Resort, Ibadan, Oyo Nigeria 2-4 May, 2017

Adams, R. (2022). 'AI in Africa: Key Concerns and Policy Considerations for the Future of the Continent. https://afripoli.org/ai-in-africa-key-concerns-and-policy-considerations-for-the-future-of-the-continent

Adomako, K., Mohamed, N., Garba, A., & Saint, M. (2018, March). Assessing Cybersecurity Policy Effectiveness in Africa via a Cybersecurity Liability Index. TPRC.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

Andreoni, A., & Tregenna, F. (2020). Escaping the middle-income technology trap: A comparative analysis of industrial policies in China, Brazil and South Africa. Structural Change and Economic Dynamics.

Asongu, S. A., Efobi, U. R., Tanankem, B. V., & Osabuohien, E. S. (2020). Globalisation and female economic participation in Sub-Saharan Africa. Gender Issues, 37(1), 61-89.

Austin, G. (2010). African economic development and colonial legacies (Vol. 1, No. 1, pp. 11-32). Institut de hautes études internationales et du développement.

Balogun, J. A. (2020). Commentary: Lessons from the USA delayed response to the COVID-19 pandemic. African journal of reproductive health, 24(1), 14-21.

Bangura, A. K. (2019). The "Arab Spring": An Epitome of Western Political Machinations.

Baylon, C., & Antwi-Boasiako, A. (2016). CHAPTER SIX: INCREASING INTERNET

Barabanov, O. N., & Maslova, E. A. (2019). The Concept of Global Commons as a Factor of Global Instability. Mirovaia ekonomika i mezhdunarodnye otnosheniia, 63(8), 55-63.CONNECTIVITY WHILE COMBATTING. Cyber Security in a Volatile World, 77.

Bello, Ganiyu, et al. "Biology Education and Bio Entrepreneur Opportunities in Nigeria." Nigerian Online Journal of Educational Sciences and Technology 1.2 (2020): 1-17.

Boo, H. W. (2017). An assessment of North Korean cyber threats. The Journal of East Asian Affairs, 97-117.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. Trends in Cognitive Sciences, 23(5), 408–422. https://doi.org/10.1016/J.TICS.2019.02.006

Bredtmann, J., Martínez Flores, F., & Otten, S. (2019). Remittances and the brain drain: Evidence from microdata for Sub-Saharan Africa. The Journal of Development Studies, 55(7), 1455-1476.

Brown, S., Korea, -Suny, & Korea, S. (2021). ARTIFICIAL INTELLIGENCE IN THE GLOBAL SOUTH (AI4D): POTENTIAL AND RISKS. https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

Castro, D., & New, J. (2016). The promise of artificial intelligence. Center for Data Innovation, 1-48.

Cherif, R., & Hasanov, F. (2019). The Leap of the Tiger: Escaping the Middle-income Trap to the Technological Frontier. Global Policy, 10(4), 497-511.

Chukwu, M. O. (2020). Projects of Economic and Social Development in the Global South: The 20th and 21st-century developmental trends and their impacts.

Corlett, R.T., Primack, R.B., Devictor, V., Maas, B., Goswami, V.R., Bates, A.E., Koh, L.P., Regan, T.J., Loyola, R., Pakeman, R.J. and Cumming, G.S., 2020. Impacts of the coronavirus pandemic on biodiversity conservation. *Biological conservation*, *246*, p.108571.

Crosston, M. D. (2011). World gone cyber MAD: How "mutually assured debilitation" is the best hope for cyber deterrence. Strategic studies quarterly, 5(1), 100-116.

Cumming, G. S. (2020). Impacts of the coronavirus pandemic on biodive ity conservation. Biological Conservation, 246, 108571.
CUNNINGHAM, M. A. (2020). A National Strategy for Synthetic Biology. Strategic Studies Quarterly, 14(3).
Deeks, A., 2020. Defend forward and cyber countermeasures. *Ashley Deeks, Defend Forward and Cyber Countermeasures, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper*, (2004).
Duncan, S. E., Reinhard, R., Williams, R. C., Ramsey, F., Thomason, W., Lee, K., ... & Murch, R. (2019). Cyberbiosecurity: A new perspective on protecting us food and agricultural system. Frontiers in bioengineering and biotechnology, 7, 63.
Dunham, K. (2008). Mobile malware attacks and defense. Syngress.
Dyer, O. (2020). Trump claims public health warnings on covid-19 are a conspiracy against him. bmj, 368, m941.
Edwards, B., & Revill, J. (2016). Decade of Synthetic Biology in the Context of the Biological and Toxin Weapon Convention and the Chemical Weapon Convention.
Endong, F. P. C., & Obi, P. (2020). "We Are Anything but a 'Shithole'Country!": Exploring Nigerian Online Journalists' Reception of Donald Trump's Insult Politics. In Deconstructing Images of the Global South Through Media Representations and Communication (pp. 273-294). IGI Global.
Englebert, P. (1997). Feature review The contemporary African state: Neither African nor state. Third World Quarterly, 18(4), 767-776.
Fa, J.E., Olivero, J., Farfán, M.Á., Márquez, A.L., Vargas, J.M., Real, R. and Nasi, R., 2014. Integrating sustainable hunting in biodiversity protection in Central Africa: hot spots, weak spots, and strong spots. *PLoS One*, *9*(11), p.e112367.
Falzon, G., Lawson, C., Cheung, K. W., Vernes, K., Ballard, G. A., Fleming, P. J., ... & Meek, P. D. (2020). ClassifyMe: a field-scouting software for the identification of wildlife in camera trap images. Animals, 10(1), 58. Bioengineering and Biotechnology, 7, 51.
Gehl Sampath, P. (2021). Governing Artificial Intelligence in an Age of Inequality. Global Policy, 12(S6), 21–31. https://doi.org/10.1111/1758-5899.12940
George, A.M., 2019. The national security implications of cyberbiosecurity. *Frontiers in bioengineering and biotechnology*, *7*, p.51.
Gewald, J.B., Leliveld, A. and Peša, I. eds., 2012. *Transforming innovations in Africa: Explorative studies on appropriation in African societies* (Vol. 11). Brill.
Gohdes, A.R. (2020), Repression Technology: Internet Accessibility and State Violence. American Journal of Political Science, 64: 488-503. https://doi.org/10.1111/ajps.12509
Goodman, W., 2010. Cyber deterrence: Tougher in theory than in practice?. *Strategic Studies Quarterly*, *4*(3), pp.102-135.
Gronvall, G. K. (2019). Synthetic Biology: Biosecurity and Biosafety Implications. In Defense Against Biological Attacks (pp. 225-232). Springer, Cham.
Gruber, K. (2019). Biohackers: A growing number of amateurs join the do-it-yourself molecular biology movement outside academic laboratories. EMBO reports 20(6), e48397.
Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence: Https://Doi.Org/10.1177/0008125619864925, 61(4), 5–14. https://doi.org/10.1177/0008125619864925
Haggard, S., & Lindsay, J. R. (2015). North Korea and the Sony hack: Exporting instability through cyberspace.
Harnischfeger, J. (2019). Biafra and secessionism in Nigeria: An instrument of political bargaining. In Secessionism in African Politics (pp. 329-359). Palgrave Macmillan, Cham.
Hernandez, L. (2021). How Rwanda's AI policy helps to shape the evolving AI ecosystem. https://digicenter.rw/how-rwandas-ai-policy-helps-to-shape-the-evolving-ai-ecosystem/
Iasiello, E. (2014). Is cyber deterrence an illusory course of action?. Journal of Strategic Security, 7(1), 54-67.
Imperiale, M., Boyle, P., Carr, P. A., Densmore, D., DiEuliis, D., & Ellington, A. (2018). Biodefense in the Age of Synthetic Biology.
Kong, D. S., & Bakker, N. (2018, August). Community driven design of living technologies. In Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2 (pp. 1-3).
Kosaki, T., Lal, R., & Reyes Sánchez, L. B. (2020, May). " Soil Education Manual-Toolbox for DIY program at your classroom" by International Union of Soil Sciences (IUSS). In EGU General Assembly Conference Abstracts (p. 21359).
Lakhdhar, Y., Rekhis, S., & Boudriga, N. (2017, December). Proactive damage assessment of cyber attacks using mobile observer agents. In Proceedings of the 15th International Conference on Advances in Mobile Computing & Multimedia (pp. 29-38).
Lewis, J. A. (2014). National perceptions of cyber threats. Strategic Analysis, 38(4), 566-576.
Marchant, E., & Stremlau, N. (2020). The Changing Landscape of Internet Shutdown in Africa| A Spectrum of Shutdowns: Reframing Internet Shutdowns From Africa. International Journal of Communication, 14, 18.
Makanda, K., Vallent, T. F., & Kim, H. (2017) Remarks on National Cyber Security for under Developed and Developing Countries: focused on Malawi.
Mas, I., & Radcliffe, D. (2010). Mobile payments go viral: M-PESA in Kenya.May, J. F., & Rotenberg, S. (2020). A Call for Better Integrated Policies to Accelerate the Fertility Decline in Sub-Saharan Africa. Studies in Family Planning, 51(2), 193-204.
MILLETT, K. K., dos Santos, E., & MILLETT, P. D. (2019). Cyber-Biosecurity Risk Perceptions in the Biotech Sector.
Ministry of Information Technology and Communication, 2017. ICT Sector Strategic Plan (2018-2024) - "Towards digital-enabled economy". Rwanda: MITEC.

Mueller, S. (2020). Facing the 2020 Pandemic: What does Cyberbiosecurity want us to know to safeguard the future?. Biosafety and Health.

Murch, R., & DiEuliis, D. (Eds.). (2019). Mapping the cyberbiosecurity enterprise. Frontiers Media SA.

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.

Ndayisaba, A. (2020). Rwanda-Burundi: Political Dialogue as a Method of Achieving Agreement. RUDN Journal of Political Science, 22(1), 105-115.

Nkusi, Fred. "Creating the National Cyber Security Authority Is Vital," April 30, 2017. https://www.newtimes.co.rw/section/read/211603.

Olubela, M. O. (2018). Shithole Countries: An Analysis of News Coverage in the US. Republic of Rwanda, National Cyber Security Strategy (Republic of Rwanda, Kigali, 2015)

Palmer, X. L., & Karahan, S. (2020, March). Defending Forward: An Exploration through the Lens of Biocybersecurity. In ICCWS 2020 15th International Conference on Cyber Warfare and Security (p. 373). Academic Conferences and publishing limited.

Palmer, X., Potter, L.N. and Karahan, S., 2021. COVID-19 and biocybersecurity's increasing role on defending forward. *International Journal of Cyber Warfare and Terrorism (IJCWT)*, *11*(3), pp.15-29.

Potter, L., Ayala, O., & Palmer, X. L. (2020). Biocybersecurity--A Converging Threat as an Auxiliary to War. arXiv preprint arXiv:2010.00624.

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.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv. https://doi.org/10.48550/arXiv.2204.06125

Richardson, L. C., Connell, N. D., Lewis, S. M., Pauwels, E., & Murch, R. S. (2019). Cyberbiosecurity: a call for cooperation in a new threat landscape. Frontiers in bioengineering and biotechnology, 7, 99.

Republic of Rwanda, National Cyber Security Strategy (Republic of Rwanda, Kigali, 2015)

Sarpong, D., Ofosu, G., Botchie, D., & Clear, F. (2020). Do-it-yourself (DiY) science: The proliferation, relevance and concerns. Technological Forecasting and Social Change, 158, 120127.

Samori, I. A., Palmer, X. L., Potter, L., & Karahan, S. (2023). Commentary on Biological Assets Cataloging and AI in the Global South. In Proceedings of SAI Intelligent Systems Conference (pp. 734-744). Springer, Cham.

Smeets, M. (2020). US cyber strategy of persistent engagement & defend forward: implications for the alliance and intelligence collection. Intelligence and National Security, 35(3), 444-453.

Tor, U. (2017). 'Cumulative deterrence'as a new paradigm for cyber deterrence. Journal of Strategic Studies, 40(1-2), 92-117.

Turianskyi, Y. (2020). Africa and Europe: Cyber Governance Lessons. South African Institute of International Affairs. https://media.africaportal.org/documents/Policy-Insights-77- turianskyi.pdf

Turner, G. (2019, May). The Growing Need for Cyberbiosecurity. In InSITE 2019: Informing Science+ IT Education Conferences: Jerusalem (pp. 207-215).

Tylecote, A. (2019). Biotechnology as a new techno-economic paradigm that will help drive the world economy and mitigate climate change. Research Policy, 48(4), 858-868.

Vinatzer, B. A., Heath, L. S., Almohri, H. M., Stulberg, M. J., Lowe, C., & Li, S. (2019). Cyberbiosecurity Challenges of Pathogen Genome Databases. Frontiers in bioengineering and biotechnology, 7, 106.

Wamuchiru, E., & Moulaert, F. (2018). Thinking through ALMOLIN: the community bio-centre approach in water and sewerage service provision in Nairobi's informal settlements. Journal of Environmental Planning and Management, 61(12), 2166-2185.

Williams, Q. (2020). 19 Rejoinders from the Shithole. Language in the Trump Era: Scandals and Emergencies, 265.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., . . . Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. The Innovation, 2(4), 100179. https://doi.org/10.1016/j.xinn.2021.100179

Lesson Plan: An Interdisciplinary Approach to Teaching Cyber Warfare Concepts

Donna M. Schaeffer[1] and Patrick C. Olson[2]
[1]Marymount University, Arlington VA USA
[2]National University, San Diego, CA USA
Donna.schaeffer@marymount.edu
polson@nu.edu

Abstract: Interdisciplinary topics and fields need special attention to ensure that the breadth of the knowledge they represent are completely expressed. This is especially important and difficult to achieve in curriculum areas that tend to add new ideas on a constant basis and are sometimes interdisciplinary themselves. The topic/field cyber warfare is area this paper considers. This paper recommends and describes a lesson plan that provides the means of achieving a fully expressed examination of cyber warfare. The outcome is an articulation of the concept that will diffuse knowledge about cyber warfare. The lesson plan will be useful to any institution (particularly universities and government agencies) that need to diffuse knowledge about cyber warfare.

Keywords: Cyber Warfare; Curriculum Development

1. Introduction

Cyber warfare is an issue that has garnered attention, both in the media and in academia. For example, a news search for "cyber warfare" on the search engine Google yielded almost 300,000 hits in September 2022. The hits ranged from book reviews on the topic, articles from professional publications such as *Chief Privacy Officer* and *Modern Diplomacy,* and news sites such as Reuters. The movie franchise "Die Hard' Comes up in a Google search for "movies about cyber warfare." Undoubtedly, the term cyber warfare has entered popular parlance.

In the academic world, a search for "cyber warfare" on scholar.google.com in September 2022 identified over 100,000 items found. The sources ranged from books to well-regarded peer reviewed journals and conference proceedings from a wide variety of disciplines including computing and engineering, diplomacy and international relations, military sciences, and law.

If we accept the definition of interdisciplinary as relating to more than one branch of knowledge (Oxford University, 2020), the justification of cybersecurity as an interdisciplinary field (Schaeffer, Olson, & Knott Eck, 2017; Singer & Friedman, 2014) may be extended to cyber warfare.

In this paper, we present a lesson plan for introducing cyber warfare concepts to a variety of disciplines and using various pedagogies. The following sections of the paper identify competencies, describe learning objectives, discuss the need for varied resources, recommend appropriate reading and multi-media resources, describe classroom activities, and provide deliverables and assessment tools. We conclude with a general discussion of introducing the concepts of cyber warfare in an interdisciplinary fashion at various levels of study.

Singer and Freidman (2014) define cyber warfare as the use of technology to attack a nation, causing harm that is comparable to the harm sustained in actual warfare. In recent years, the definition has morphed into "actions by a nation-state or international organization to attack and attempt to damage another nation's computers or information networks through such as activities as computer viruses or denial-of-service attacks (Rand, 2020)." Gazula (2017) asserts "To many, cyber-warfare represents the 5th battle-space—a new type of warfare in need of further definition. To others, it is merely a new weapon to be integrated into traditional conflict."

The discourse on cyber warfare varies among disciplines, from a focus on nation-states in political science and international relations, terrorism in criminal justice studies, to protect and defend in computing. Table 1 displays excerpts from course descriptions in several disciplines. These course descriptions provide an overview of course-level inclusion of cyber warfare in programs. We have used these and other course descriptions to distill important concepts for one learning module about cyber warfare within a related course e.g., cybersecurity, criminal justice, and international relations.

Table 1: Excerpts from Cyberwarfare Course Descriptions

Excerpt	Institution	Discipline	Date
This course provides an overview of cyber warfare and the potential impact of its use by military, terrorist, and criminal organizations. An overview of cyber weaponry will be presented, and various offensive and defensive strategies will be examined via case studies.	American Public Universities	International Relations	nd
This course examines legal and policy challenges stemming from rapidly evolving cybersecurity threats. The course will begin with an introduction to cyber security and cyber warfare, and then explore the national and international legal frameworks that govern the cyberspace, including laws related to cybercrime, cyberespionage, and cyberwar.	Rutgers University	Law	nd
Provides an introduction to counterintelligence, with a focus on the evolution of contemporary counterintelligence in military, government, and pseudo-government organizations, both domestically and internationally. The course will also address terrorism as a criminally violent tactic used to achieve political or social goals and will examine individuals and groups, their motives and tactics, and how government and law enforcement have responded through investigation, prosecution, and punishment.	Marymount University	Criminal Justice	2020
This course examines traditional theories, concepts, and practices in international relations and warfare- conventional, unconventional, and modern- and relates them to the emerging dynamics of cyber war . . . The principal objective of this course is to introduce students to cyber war within the context of traditional, and emerging, concepts of armed and unarmed warfare.	Carnegie Mellon University	Political Science	2020
This course explores cyber warfare from an electrical engineering perspective. Rudimentary denial-of service techniques through intelligent wave form specific forms of computer network attack (CNA) are covered.	Naval Postgraduate School	Electrical Engineering	nd

2. Competencies

Learning competencies are skills students should acquire in a lesson, course, or degree program. Competencies may be identified by professional associations, disciplinary agreement, accrediting bodies, or by individual instructors. As cyber warfare is a relatively new academic topic, competencies from existing courses may introduce ideas for developing basic competencies we could expect learners to achieve from one lesson plan on cyberwarfare that is embedded within a wide range of disciplinary courses. Our suggestions are shown in Table 2.

Table 2: Suggested Competencies

After successfully completing this learning module, students will:
• Be familiar with the historical emergence and evolution of cyber warfare
• Be introduced to global perspectives and policies on cyber warfare
• Be acquainted with cyber warfare capabilities of a selected nation-state

3. Reading Materials.

To prepare lecture materials, instructors should read the 65-page tome from the Army War College entitled *Cyber Defense: An International Review* (Giles & Hartmann, 2015). This book provides a survey of approaches by the German, Swedish, Norwegian, and Estonian defense organizations. They should also read *The Emerging Risk of Virtual Societal Warfare* (Mazarr, Bauer, Casey, Heintz, & Matthews, 2019), a recent report from Rand Corporation.

College-level reading for students involves not only being able to read the text, but also to draw conclusions from it. Comprehension is a skill that students can develop. A course module on a sensitive topic, such as cyber warfare, requires student to read critically. Recommended readings for students in preparation for the lesson include short articles and book excerpts from media and think tanks across the political spectrum. Suggestions include (current URLs for each resource are provided in the references):

- The Wired Guide to Cyberwar by Andy Greenburg, 23 September 2019.
- "Writing the Rules of Cyberwar" by Alyza Sebenius in The Atlantic. 28 June 2017.
- Our Adversaries are Using Cyberwarfare. We Must Be Prepared by James Di Pane and Alexandra Marotta. 29 July 2019.
- Waging (cyber)war in Peacetime by Fergus Hanson. 22 October 2015.

4. Multimedia Resources.

Multimedia resources that could be employed either in a "flipped classroom" approach where students watch the video or listen to the audio before a class session or shown as an introduction to a class period. We recommend a brief 10-minute TedTalk by Rodrigo Bijou called "Governments Don't Understand Cyberwarfare. We Need Hackers." A current URL is provided in the References.

Several media outlets have posted recent short videos on cyberwarfare, including business outlets such as *Forbes* and *Fortune,* political outlets including C-Span and *Politico*, and general interest publications such as *National Geographic*. A professor needs to search videos on a search engine and select one from a trusted source or several that represent different orientations. Selection, and subsequent class activities, should be grounded in the cognitive theory of multimedia learning (Mayer and Moreno, 2003). This theory advances that students use multiple channels, that is visual and auditory, to make sense of new information. Figure 1 represents how information moves from a resource to students' knowledge base.

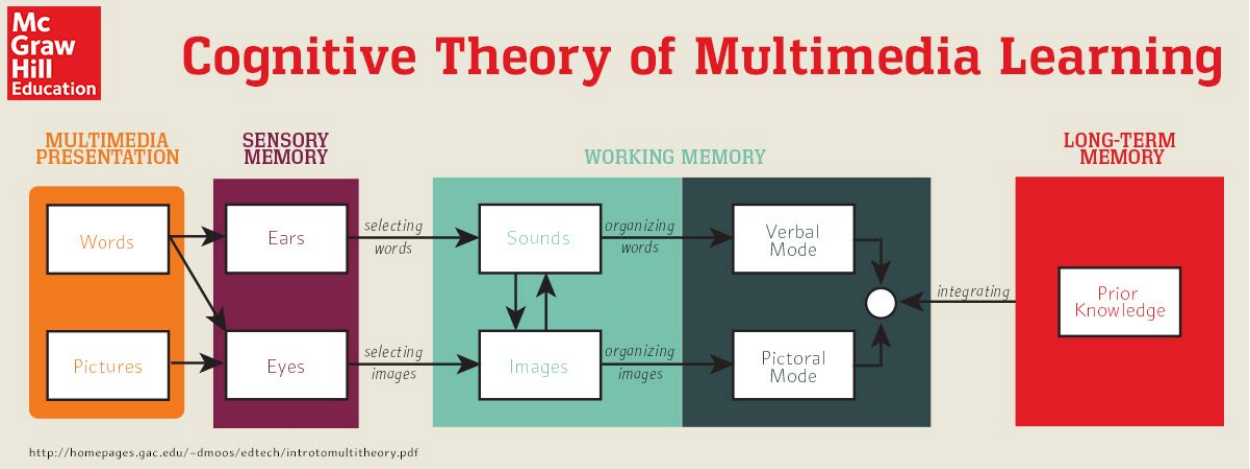


Figure 1: Cognitive Theory of Multimedia Learning (Source: https://www.mheducation.ca/blog/richard-mayers-cognitive-theory-of-multimedia-learning)

The Cognitive Theory of Multimedia Learning (CTML) recognizes that humans have a limited capacity for information. That is why we suggest presenting the same information in multiple ways to reinforce a limited number of cyberwarfare concepts, rather than a broad scope. It is also a rationale for selecting short videos.

5. Classroom Activities.

The CTML suggests learners actively engage with the material in order to learn it (McGrawHill Canada, 2019). After a brief lecture based on the suggested reading materials for instructors, there are several interactive activities that could be completed in the classroom or via online means in learning management systems. Activities should be designed around having students model cause-and-effects, make comparisons, apply generalizations, create enumerations, or develop classifications (Mayer and Moreno, 2003). We suggest activities and topics in the following section of this paper.

Reflection paper assignments enable students to think about how the learning experience changed them and the insights they have gained. Yancy (1998) describes reflection as needing both review of what has been learned and a projection of the student may use the information in the future.

Author F. Scott Fitzgerald is attributed with saying "The test of a first-rate intelligence is the ability to hold two opposed ideas in mind at the same time and still retain the ability to function." This describes debate assignments. Debates force students to see more than one side of an issue. Since cyberwarfare is a sensitive topic, and one on which students may come into the classroom with pre-conceived ideas, debates could add value. Debates also help build communication, collaboration, and critical thinking skills.

Role-playing is a pedagogical technique that is appropriate for complex concepts (Sogunro, 2004). Westrup and Planander (2013) found that role-playing increases both knowledge retention and student engagement. It helps student practice empathy and seeing issues from various perspectives (Elmore, n.d.).

Table 3 provides sample assignment topics that may be appropriate for the pedagogies discussed in the paper.

Table 3: Pedagogies and Appropriate Assignment Topics

Reflection Paper	Does cyber warfare differ from traditional warfare? If yes, how? If not, why not? (Sciarrone, 2017)
Debate	Statement: International interference in national elections is an act of war.
Role-playing	A group of government leaders have met to discuss what actions may be taken to prevent cyber warfare.

6. Conclusion.

This paper began by noting that cybersecurity and therefore cyber warfare is interdisciplinary and thus entire courses on cyber warfare are taught at various levels of study and in different disciplines are taught. To support the cyber warfare topic a lesson plan for a brief class section on cyber warfare that can be adapted to all levels of study, as well as tailored to specific disciplines was presented. The discussion included excerpts from courses, identification of competencies, recommended reading material and classroom activities.

References

Bijou, R. (2015, June). *Governments don't understand cyber warfare. We need hackers.* Retrieved from ted.com: https://www.ted.com/talks/rodrigo_bijou_governments_don_t_understand_cyber_warfare_we_need_hackers

Di Pane, J., & Marotta, A. (2019, July 29). *Our Adversaries Are Using Cyberwarfare. We Must Be Prepared.* Retrieved from heritage.org: https://www.heritage.org/cybersecurity/commentary/our-adversaries-are-using-cyberwarfare-we-must-be-prepared

Elmore, L.B. (n.d.) *Role-playing.* Retrieved from https://ablconnect.harvard.edu/role-play-research

Gazula, M. B. (2017). *Cyber Warfare Conflict Analysis and Case Studies.* MIT Management Sloan School, (IC)3. Cambridge, MA: Massachusetts Institute of Technology. Retrieved 7 24, 2020, from http://web.mit.edu/smadnick/www/wp/2017-10.pdf

Giles, K., & Hartmann, K. (2015). *Cyber Defense: An International View.* Carlisle Barracks, PA: United States Army War College Press. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a622264.pdf

Greenburg, A. (2019, 8 23). *The WiRED Guide to Cyberwar.* Retrieved from wired.com: https://www.wired.com/story/cyberwar-guide/

Hanson, F. (2015, October 22). *Waging (cyber)war in peacetime.* Retrieved from brookings.edu: https://www.brookings.edu/blog/up-front/2015/10/22/waging-cyberwar-in-peacetime/

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

Mazarr, M. J., Bauer, R. M., Casey, A., Heintz, S. A., & Matthews, L. J. (2019). *The Emerging Risk of Virtual Societal Warfare; Social Manipulation in a Changing Information Environment.* Santa Monica, CA: Rand. Retrieved from https://www.rand.org/pubs/research_reports/RR2714.html

McGrawHill Canada (2019, 16 April). *Richard Mayer's Cognitive Theory of Multimedia Learning.* Retrieved from McGrawHill Canada: https://www.mheducation.ca/blog/richard-mayers-cognitive-theory-of-multimedia-learning

Oxford University. (2020, 7 24). *interdisciplinary.* Retrieved from lexico.com: https://www.lexico.com/en/definition/interdisciplinary

Rand. (2020, 7 24). *Cyber Warfare.* Retrieved from rand.org: https://www.rand.org/topics/cyber-warfare.html#:~:text=Cyber%20warfare%20involves%20the%20actions,denial%2Dof%2Dservice%20attacks.

Schaeffer, D. M., Olson, P. C., & Knott Eck, C. (2017). An Interdisciplinary Approach to Cybersecurity Curriculum. *Journal of Higher Education Theory and Practice, 17*(9), 36-40.

Sciarrone, M. (2017). Cyber Warfare: The New Front. *The Catalyst.* Issue 6. Retrieved from the George W. Bush Institute https://www.bushcenter.org/catalyst/modern-military/sciarrone-cyber-warfware.html#:~:text=Cyber%20warfare%20could%20make%20conventional,little%20use%20in%20the%20future.

Sebenius, A. (2017, June 28). *Writing the Rules of Cyberwar.* Retrieved from theatlantic.com: https://www.theatlantic.com/international/archive/2017/06/cyberattack-russia-ukraine-hack/531957/

Singer, P. W., & Friedman, A. (2014). *Cybersecurity and Cyberwar: What Everyone Needs to Know.* Oxford: Oxford University Press.

Sogunro, O.A. (2004). Efficacy of role-playing pedagogy in training leaders: Some reflections. *Journal of Management Development*, 23(4), 355-371

Yancy, K.B. (1998). *Reflection in the Writing Classroom.* Retrieved from Utah State University: https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1119&context=usupress_pubs&_ga=2.227644339.927502929.1663441304-1372834558.1663441304

Westrup, U. & Planander, A. (2013). Role-play as a pedagogical method to prepare students for practice: The students' voice. Ogre utbildning, 3(3), 199-210.

The Impact of Edge Computing on the Industrial Internet of Things

Nkata Sekonya[1] and Siphesihle Sithungu[2]
University of Johannesburg, Johannesburg, South Africa

nkatasekonya61@gmail.com
siphesihles@uj.ac.za

Abstract: The emergence of the Industrial Internet of Things (IIoT) has improved the management of industrial operations and processes. IIoT involves collecting and processing data from a vast array of sensors deployed across industrial complexes. This enables the measurement of the efficiency of industrial processes, monitoring the health of machinery, optimisation of operations, and response to real-time events. In its application, IIoT underpins the operation of critical infrastructure in sectors including manufacturing and utilities. Maintaining the availability and resilience of critical infrastructure against internal and external threats is essential to minimise disruptions that could have a debilitating effect on a nation's economy. Although internal threats can lead to a critical infrastructure's downtime, external threats through cyberattacks also pose a significant threat. Historical events have demonstrated that the successful disruption of critical infrastructure can lead to the loss of human life, the interruption of necessary economic activities and national security concerns. Therefore, the availability of resilient critical infrastructures is vital to the well-being of a country. In this context, the paper compares the deployment of traditional IIoT to that of edge computing for the storage and processing of data. Traditional IIoT relies on a centralised server for data storage and processing, which is insufficient as IIoT environments cannot tolerate delays in responding to real-time events. Conversely, edge computing allows for data processing at the edge, closer to the data source, which plays a crucial role in enabling IIoT devices to respond to real-time events by reducing decision-making latency. Moreover, the decentralised nature of edge computing reduces the reliance on a centralised server by only sending required data to the cloud for further processing. Although edge computing enhances IIoT deployments, a notable concern is a resultant increase in the attack surface of IIoT environments, which consequently restricts its implementation. Exploratory research is conducted to explore the integration of edge computing into IIoT environments with a focus on improving the management and operation of critical infrastructures. A review of the current literature is performed to identify and discuss security concerns prevalent in edge computing-enabled IIoT environments and proposed mitigation strategies.

Keywords: Industrial Internet Of Things, Edge Computing, Critical Infrastructure, Cybersecurity

1. Introduction

The "Internet of Things" (IoT) refers to a network of devices with sensors and actuators that use the internet as a communication medium to gather, process, and transfer data (Koohang et al., 2022). IoT has gained popularity in human-centric applications since its inception, including connected homes, smart watches, and productivity software. As IoT expanded from human-centric applications to control industrial processes, the "Industrial Internet of Things" (IIoT) came into existence. A subset of IoT called IIoT is focused on the security and efficiency of industrial processes (Bhaiyat and Sithungu, 2022). The development of the IIoT has the potential to enhance critical infrastructure sectors, including manufacturing and utilities' management and operation. The IIoT establishes data gathering and analysis from a variety of sensor-equipped industrial equipment, enabling better industrial process efficiency, data-driven decision-making, monitoring the health of industrial machinery and optimising operations (Qiu et al., 2020). Since critical infrastructure sectors cannot tolerate delays in responding to real-time events, traditional IIoT setups that rely on a central server for storage and processing are insufficient (Qiu et al., 2022). By processing the data closer to the source, edge computing lowers the latency associated with data transfer to a centralised server (Mirani et al., 2022). However, edge computing is not without its critical flaws. The expanded attack surface that edge computing introduces to IIoT systems is a significant issue that limits its incorporation into IIoT environments. The objective of this paper is to conduct a literature review on the current state of edge computing in IIoT environments, with an emphasis on enhancing the management and operation of critical infrastructures. Security concerns common in edge computing-enabled IIoT environments are identified and discussed, together with proposed mitigation strategies.

To accomplish this objective, the paper is structured as follows: Section 2 discusses cloud, fog, and edge computing in the context of IIoT. Section 3 delves into the integration of edge computing into IIoT environments. Section 4 discusses the role of edge computing in improving critical infrastructure management and operation. Section 5 discusses the security concerns prevalent in edge-computing-enabled IIoT environments, as well as approaches to securing these environments. Section 6 concludes the paper.

2. Cloud, Fog, and Edge Computing

The "Industrial Internet of Things" (IIoT) establishes the collection, communication, and processing of data from a vast array of sensors deployed across industrial complexes (Goswami, Jadav and Soni, 2022). IIoT is an extension of the "Internet of Things", focusing on the industrial applications of IoT. IIoT acts as a catalyst to enhance the management of industrial processes and operations, allowing for monitoring of the health of machinery, optimisation of operations, and response to real-time events. The Industrial Internet Reference Architecture (IIRA) is an open architecture for the industrial internet (Pivoto et al., 2021). The IIRA presents a functional view of an IIoT deployment typical across industries. The functional domains and crosscutting functions are concerned with the functionality of a system, while the system requirements are concerned with efficiency, i.e., how well the system works (Industrial Internet Consortium, 2019). Figure 1 depicts the functional domains, crosscutting functions, and system requirements of IIoT environments.

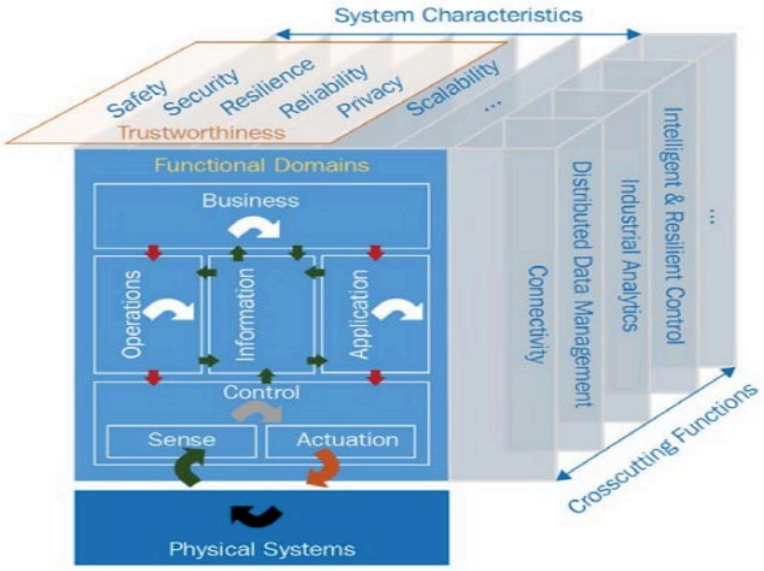


Figure 1: IIRA functional domains, crosscutting functions, and system requirements (Industrial Internet Consortium, 2019)

The requirements of IIoT systems include scalability, efficient storage management, fault tolerance, security, computational proximity, data transfer, latency minimisation, and safety assurances (Industrial Internet Consortium, 2019; Chalapathi et al., 2021; Zhang, Li and Chen, 2021; Sittón-Candanedo et al., 2019). Computing infrastructure is necessary for IIoT environments to store and process data that has been acquired. There are also several computing paradigms to be explored in relation to IIoT, such as cloud, fog, and edge computing. For IIoT contexts, some computer paradigms are more efficient than others. As such, the remainder of this section reviews cloud, fog, and edge computing.

2.1 Cloud Computing

Cloud computing refers to the practice of leasing computer resources from a centralised pool of shared capacity, with consumption taking place over the Internet (Zhang, Li and Chen, 2021). Cloud computing is an enabler of IoT because it provides scalable computing resources and data storage for IoT sensors. Traditional cloud computing, on the other hand, is insufficient for IIoT environments, where delays in responding to real-time events are unacceptable. The requirements of IIoT environments include low latency, reliable communication, and security, which existing cloud infrastructure can hardly meet (Angel et al., 2021), especially as IIoT devices continue to generate large amounts of data.

Other cloud computing limitations include the location of the centralised servers, as some countries may have laws prohibiting the storage of critical data in other countries (Basir et al., 2019). Computing and storage resources in a cloud computing setup may be exhausted as all data collected by IIoT sensors is transferred to a centralised server for storage and processing (Mirani et al., 2022). To offload the burden on the servers, processes must be in place to determine the criticality of the data before processing. By bringing resources closer to the IIoT environment, fog computing aims to address challenges encountered in IIoT environments that rely on centralised servers.

2.2 Fog Computing

Fog computing is a cloud computing extension that provides a bridge from the data centre to the edge to reduce network latency (Caiza et al., 2020). A fog computing network comprises nodes closer to the IIoT environment, with each node acting as a lightweight version of a cloud computing server (Caiza et al., 2020). This is done to offload the processing burden and shorten the time it takes to process and analyse the collected data (Mirani et al., 2022). In the management of IIoT environments, fog computing outperforms traditional cloud computing. Fog computing is less centralised than cloud computing and does not have a single point of failure (Zhang, Li and Chen, 2021).

2.3 Edge Computing

Edge computing provides an advantage over cloud and fog computing by allowing data processing at the edge, reducing the resources required for cloud computing by sending only the required data to the cloud for further processing (Mirani et al., 2022). Edge computing reduces data transfer latency, allowing for quick responses to real-time events. Edge computing also reduces the computing and storage requirements of cloud servers, lowering the cost of running the cloud infrastructure significantly (Basir et al., 2019). Figure 2 depicts the various computing paradigms.

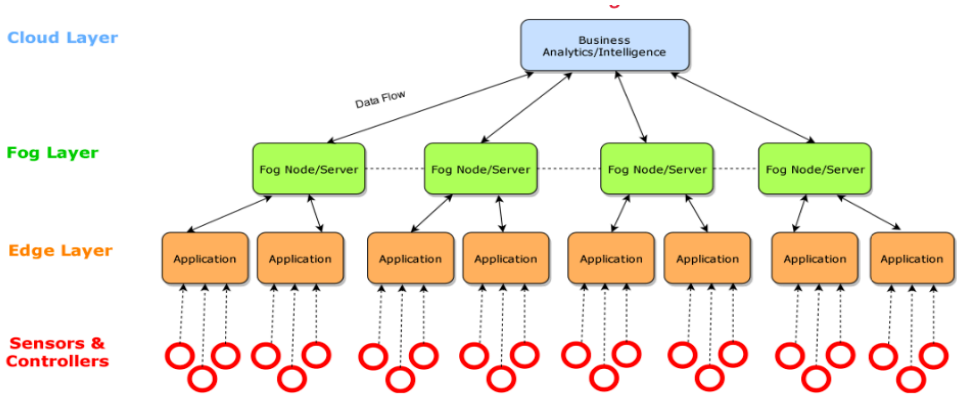


Figure 2: Industrial IoT data processing layer stack (Chalapathi et al., 2021)

The purposes of cloud, fog, and edge computing are similar in that they all aim to provide scalable computing and storage resources. The difference is in how the targets are met; cloud computing offers these resources through servers located in remote locations. By deploying lightweight servers closer to the IIoT environment, fog computing creates a bridge between the IIoT environment and the cloud. These resources are provided at the *edge*. Table 1 presents a comparison of cloud, fog, and edge computing. The following section goes over the role of edge computing in IIoT.

Table 1: Comparison between cloud, fog, and edge computing

Criteria	Cloud Computing	Fog Computing	Edge Computing
Deployment Nature	Relies on centralised servers in remote data centres for storage and computation	Acts as a bridge between remote data centres and the data source, with fog computing nodes acting as a lightweight version of a cloud computing server	Deployed at the network edge, closer to the data source
Computing and Storage Capacity	Scalable	Fairly Scalable	Limited
Advantages	Provides scalable computing resources and data storage	Offers improved data transfer compared to cloud computing	Reduced data transfer latency, enabling prompt response to real-time events
Disadvantages	Increased data transfer latency because of the distance between the data source and remote servers	Exhibits centralisation as it relies on relatively few fog nodes	Limited computing and storage capacity

3. The Integration of Edge Computing into Industrial IoT

Traditional IIoT devices have limited resources and cannot efficiently store and process data collected in IIoT environments (Khan et al., 2020). Conventionally, this issue was addressed by having the IIoT device send data to centralised servers for storage and processing. However, delays in data transfer to centralised servers made it difficult for IIoT devices to respond to real-time events promptly (Basir et al., 2019). Edge computing was proposed as a solution for reducing data transfer latency by performing computations on the network's edge, which is close to the data source (Mirani et al., 2022). Only the necessary data is transferred to the centralised server for further processing in an edge computing environment, reducing the latency associated with data transfer (Qiu et al., 2020).

IIoT deployments must meet requirements such as data transfer latency minimisation, computational proximity, scalability, efficient storage management, fault tolerance, security, cost-effectiveness, interoperability between heterogeneous IIoT systems, and safety assurances to be a viable option for industrial applications (Industrial Internet Consortium, 2019; Chalapathi et al., 2021; Zhang, Li and Chen, 2021; Sittón-Candanedo et al., 2019). Additionally, IIoT should allow for consistent management and operation of industrial processes. IIoT environments with edge computing capabilities are scalable because increased data transfer can be addressed by adding more edge nodes. These additional edge nodes should be able to raise data transfer capacity while reducing network latency (Chalapathi et al., 2021). Edge computing also enables efficient offloading strategies for determining which processes should be computed at the network core and which at the edge (Zhang, Li and Chen, 2021). Edge computing nodes enable effective storage management by distributing data around the network's edge, increasing network resilience, and sparing storage resources by only transferring the necessary data to the centralised servers (Sittón-Candanedo et al., 2019).

Because IIoT environments are distributed, edge computing ensures fault tolerance. The failure of one edge device should not bring the entire network to a halt. In a centralised cloud computing environment, this type of fault tolerance is difficult to achieve because the disruption of a centralised server will halt network operation, rendering IIoT devices useless (Chalapathi et al., 2021). Cloud computing environments present a single point of failure compared to edge computing environments. Due to their distributed nature, edge computing devices help to ensure the security of the data collected to some extent. To gain access to any useful information, a threat actor would need to compromise many edge devices (Sittón-Candanedo et al., 2021). A compromise of one edge device may be ineffective. Data privacy is also ensured by the isolation of each network node (Qiu et al., 2020).

IIoT devices produce vast volumes of data that must be transferred (Qiu et al., 2020). Edge computing nodes communicate only the required information, minimising the quantity of data that must be stored and processed in cloud computing environments (Angel et al., 2021). The development of reference architectures to standardise the integration of edge computing into IIoT systems is receiving more attention from industry and academia (Qiu et al., 2020). Since it guarantees consistent outcomes across different edge computing implementations, this will accelerate the adoption of edge computing. By standardising the collected data into a format understood by systems in the upper tiers of the IIoT stack, edge computing nodes can promote interoperability across systems in IIoT environments (Antao et al., 2018). Edge computing nodes also enable quick problem detection in IIoT contexts, safeguarding industrial and worker safety (Angel et al., 2021).

Industries requiring reliable data collection and processing for time-sensitive projects have adopted this computing paradigm due to the performance gains brought about by the integration of edge computing into IIoT systems (Sittón-Candanedo et al., 2019). Manufacturing, mining, transportation, and healthcare are a few of these. To better deliver medical services like the monitoring of ECG data and the processing of MRI data, the healthcare industry has begun implementing IIoT technologies (Chalapathi et al., 2021). However, the majority of these solutions rely on cloud-based processing and storage, which is inconvenient for time-sensitive operations such as responding to heart attacks. Due to the benefits of reduced latency and improved reliability, there has been a gradual shift towards edge-enabled IIoT. Table 2 depicts the benefits of incorporating edge computing into IIoT environments.

Table 2: The Integration of edge computing into IIoT

IIoT Requirement	Edge Computing Enhancement
Minimisation of the latency associated with data transfer; Computational Proximity	Computations are performed on the edge of the network, closer to data sources
Scalability	Increased data transfer can be addressed by the introduction of additional edge nodes
Storage and Computational Offloading	Computations are performed at the edge; only necessary data is sent for further processing at the network core
Efficient Storage Management	Storage is distributed across the edge of the network ensuring increased resilience
Fault Tolerance	The distributed nature of edge computing ensures fault tolerance
Security	The distributed nature of edge computing environments makes it difficult to compromise
Cost Effectiveness	Only necessary data is transferred to the cloud, reducing the cost associated with storage and processing
Consistent management and operation of industrial processes	Reference architectures ensure consistent results from the deployment of edge computing
Interoperability between heterogeneous IIoT systems	Nodes can format the collected data in a manner understandable to systems in the upper layers on the IIoT stack
Safety	Allows for the timely detection of faults in IIoT environments

Although edge computing has improved the efficiency of IIoT environments, significant issues exist that impede its growth, the most important of which is security (Zeyu et al., 2020). Edge computing devices are typically placed in easily accessible areas, posing a security risk to the devices. Because these devices have limited resources and are relatively complex, traditional security mechanisms are difficult to implement. The following section discusses the security concerns prevalent in edge computing devices.

4. Security Concerns of Edge Computing in IIoT

IIoT underpins the operation and management of critical infrastructures, meaning cyberattacks targeted at IIoT environments can cause disruptions in crucial economic activities. They can lead to devastating effects on the health and safety of industrial workers (Angel et al., 2021). The integration of edge computing further increases the attack surface in IIoT environments, especially if it involves the addition of dedicated edge devices, which present difficulties in assuring security. The following subsections discuss security concerns resulting from the integration of edge computing into IIoT environments, with table 3 summarising the reviewed literature.

4.1 Internal Threats

Many contexts, including edge computing, frequently experience internal threats. Internal threats include malicious attacks such as sabotage and unintentional attacks brought on by human error or ineptitude (Gaidarski and Minchev, 2021). Employees who are not adequately trained to carry out their responsibilities securely pose a security risk to the company. In addition, disgruntled employees constitute a severe threat to the company's security. The deployment of malicious software in edge computing environments and the unintentional disclosure or deletion of sensitive data are two examples of threats in this category. Training personnel to carry out their responsibilities appropriately and securely is one measure to prevent internal hazards (Bhaiyat and Sithungu, 2022). Another measure is behaviour analysis and risk profiling (Gaidarski and Minchev, 2021).

4.2 Identification and Authentication

Mechanisms for identification and authentication focus on allowing only authorised users to access the system (Alwakeel, 2021). This involves the identification and authentication of IIoT devices in the context of edge computing and IIoT. Edge computing environments in IIoT contain many devices, making it challenging to identify rogue devices (Zhang, Li and Chen, 2021). An IIoT network can be breached using rouge devices as the first attack vector (Husain and Askar, 2021). Hardware-based authentication methods and access control monitoring are examples of mitigations that can prevent authentication-based intrusions (Chalapathi et al., 2021).

4.3 Authorization

Authorisation is concerned with access rights, which in the context of edge computing refers to the actions an edge node is permitted to perform (Zhao et al., 2021). Edge computing node access rights can pose security issues. A compromise of an edge node in an IIoT environment with overreaching access rights can result in

privilege escalation, which could grant an attacker further control of the IIoT network (Qiu et al., 2020). To protect against this type of attack, edge computing nodes should be granted only the access rights required to carry out their duties (Zhao et al., 2021).

4.4 Data Security

Vast volumes of data are produced in edge and IIoT contexts. Although this data enables data-driven decision-making, security issues with the data's storage and transmission exist. Compromises to the data's availability, confidentiality, and integrity are among the security threats (Chalapathi et al., 2021). Such compromises may result in the modification of critical data, resulting in system failures (Zhang, Li and Chen, 2021). Mitigations include employing mechanisms to perform integrity checks, using lightweight encryption mechanisms to avoid sensitive data disclosure, and implementing Denial of Service (DoS) protections (Parikh et al., 2019).

4.5 Network Attacks

Edge computing employs a variety of communication networks, including wired and wireless networks, as well as the internet (Roman, Lopez and Mambo, 2018). These communication networks introduce security risks that could potentially disrupt critical infrastructure operations. Denial of Service (DoS) attacks, which prevent legitimate users from accessing network resources, and Man in the Middle attacks, which can result in information disclosure, modification, and destruction, are examples of attacks in this category (Roman, Lopez and Mambo, 2018). Mitigation strategies include using Intrusion Detection and Prevention Systems (IDPS) to detect and prevent network attacks (Parikh et al., 2019). Furthermore, systems in the edge computing environment should be isolated to avoid lateral movement by threat actors (Bhaiyat and Sithungu, 2022).

4.6 Side Channel Attacks

Side Channel attacks attempt to exploit a system by gathering publicly available information to deduce sensitive private information (Patil et al., 2020). Communication protocols and embedded sensors are popular side channels in edge computing systems (Xiao et al., 2019). The systems may leak some operational data, which threat actors may use to deduce sensitive information. Side channel attacks can be mitigated by randomising publicly available data to avoid targeted attacks and restricting side channels altogether (Xiao et al., 2019).

Table 3: Security concerns and countermeasures

Security Issue	Examples	Countermeasures	References
Internal Threats	Accidental disclosure or deletion of critical data, deployment of malicious software	Employee training, Behaviour analysis and risk profiling	(Bhaiyat and Sithungu, 2022); (Gaidarski and Minchev, 2021);
Identification and Authentication	Rogue IIoT devices	The implementation of hardware-based authentication schemes, monitoring of access control attempts	(Zhang, Li and Chen, 2021); (Husain and Askar, 2021); (Chalapathi et al., 2021)
Authorisation	Privilege escalation	Limited access rights	(Qiu et al., 2020); (Zhao et al., 2021)
Data Security	Compromises to the confidentiality, integrity, and availability of data in storage or transmission	The implementation of integrity validation and encryption mechanisms, Denial of Service protection	(Parikh et al., 2019); (Chalapathi et al., 2021); (Cao et al., 2020)
Side Channel Attacks	Inference of sensitive information	Restrictions of side channels, randomisation of publicly available data to avoid targeted attacks	(Xiao et al., 2019); (Patil et al., 2020)
Network Attacks	Denial of Service (DoS) and man-in-the-middle attacks	Network segmentation and the deployment of Intrusion Detection and Prevention systems	(Roman, Lopez and Mambo, 2018); (Husain and Askar, 2021); (Parikh et al., 2019)
Legacy Systems	Known attacks on outdated systems	Change and vulnerability management	(Patil et al., 2020)
Physical Security	Physical damage and tampering with the device	Physical security measures such as guards, barriers, strict access controls and CCTV cameras	(Roman, Lopez and Mambo, 2018)
Digital Forensics Readiness	Poor effectiveness and efficiency of digital forensic investigations	Establish mechanisms to collect and perverse potential digital evidence	(Kebande, Karie and Venter, 2018); (Stoyanova et al., 2020); (Shalaginov, Iqbal and Olegård, 2020);

4.7 Legacy Systems

Edge computing is essential to operating critical infrastructures that cannot afford downtime. Furthermore, edge computing is used in environments that run legacy systems not built with security in mind (Bhaiyat and Sithungu, 2022). Edge nodes are thus vulnerable to cyber-attacks. Other issues that arise as a result of the use of legacy systems include the management and deployment of updates, which include security updates (Patil et al., 2020). Change and vulnerability management should be employed to better prepare for attacks and manage vulnerabilities of legacy systems.

4.8 Physical Security

By bringing edge nodes closer to IIoT devices, edge computing lowered the latency associated with data transfer to centralised cloud servers, raising questions about physical security. Threat actors can physically harm edge nodes and IIoT devices if there are no physical security measures in place (Khan et al., 2020). Critical infrastructure operations may be disrupted as a result of the physical damage, endangering human life. Security guards, physical barriers that restrict access to IIoT locations, and CCTV cameras are examples of mitigations against physical damage (Roman, Lopez and Mambo, 2018).

4.9 Digital Forensics Readiness

Edge computing integration into IIoT environments introduces new challenges to the field of digital forensics (Stoyanova et al., 2020). The vast amounts of data produced in IIoT environments can aid digital forensics investigations. However, the variety of devices, distributed nature, and non-standard formats present challenges (Stoyanova et al., 2020). Previously, (Kebande, Karie and Venter, 2018) proposed an IoT architecture incorporating Digital Forensics Readiness (DFR) as a security component in IoT environments. DFR ensures the efficiency and effectiveness of digital forensics investigations by establishing mechanisms for gathering and preserving evidence in the event of a security incident (Kebande, Karie and Venter, 2018). Digital forensics readiness in edge computing environments is still in its early stages. (Shalaginov, Iqbal and Olegård, 2020) present a methodology for digital forensics readiness in edge computing environments in their research.

5. How Edge Computing can transform critical infrastructure sectors

The management and operation of critical infrastructures may change as a result of edge computing-enabled IIoT. This is significant since a nation's economic health depends on its critical infrastructure (Lanz, 2022). Because they are distributed, edge computing systems are fault resistant; a disruption on one node should not affect others. For critical infrastructures that cannot afford downtimes, this is transformational. Critical infrastructure sectors can potentially become more resilient to attacks and disruptions when the advantages provided by edge computing are appropriately leveraged.

Edge computing limits the amount of data it delivers to the central server, which makes it possible to run critical infrastructures effectively because operators are not overloaded. The data transfer mechanism also lowers data transfer latency, enabling operators to get crucial data promptly and make data-driven choices. Critical infrastructure systems and equipment tend to be distributed and complicated, making maintenance and troubleshooting challenging. Operators may be able to remotely check on the condition of these components thanks to IIoT and edge computing devices. However, this creates security issues because threat actors may exploit the same tools for maintenance for malicious reasons.

Finally, the convergence of Information Technology (IT) and Operational Technology (OT) has increased in critical infrastructure sectors (Angel, 2021). Historically, IIoT environments in critical infrastructure sectors like utilities were disconnected from the internet (Murray, Johnstone and Valli, 2017). The increased efficiencies brought about by the growth of the internet necessitated the integration of traditional OT and IT environments. Because OT systems were never designed with security in mind, the convergence introduces security risks. Edge computing can help alleviate some security concerns by implementing security controls at the network's edge to ensure secure data storage and transmission.

6. Conclusion

The paper's objective was to address edge computing's incorporation into IIoT contexts, with a particular emphasis on the management of crucial infrastructures. In support of this, the paper pointed out that IIoT includes IoT's industrial applications. The paper also contrasted the application of edge and cloud computing in

IIoT contexts. Edge computing is more suited for IIoT contexts because it lowers the latency associated with data transfer to centralised servers.

The integration of edge computing into IIoT was then covered in this paper, with advances in safety, security, and the efficient use of storage and computing resources noted. Edge computing makes it possible to manage critical infrastructures consistently. There was a discussion of the widespread security issues in edge computing environments. Internal dangers brought on by uninformed or disgruntled personnel, attacks on authentication and authorisation mechanisms, data security, side-channel attacks, network attacks, legacy systems, and physical security were all discussed.

Finally, the paper briefly explored IoT Digital Forensics Readiness (DFR), a new research area focusing on securing edge computing environments against potential threats. DFR sets up procedures that enable a quick reaction to a security incident. This paper's objective was thus accomplished. Future work should investigate how edge computing has impacted the convergence of Information and Operational Technology in IIoT environments.

References

Angel, N.A., Ravindran, D., Vincent, P.D.R., Srinivasan, K. and Hu, Y.C. (2021) Recent advances in evolving computing paradigms: Cloud, edge, and fog technologies. *Sensors*, 22(1), p.196.

Antao, L., Pinto, R., Reis, J. and Gonçalves, G. (2018, April) Requirements for testing and validating the industrial internet of things. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 110-115). IEEE.

Alwakeel, A.M. (2021) An overview of fog computing and edge computing security and privacy issues. *Sensors*, 21(24), p.8226.

Basir, R., Qaisar, S., Ali, M., Aldwairi, M., Ashraf, M.I., Mahmood, A. and Gidlund, M. (2019) Fog computing enabling industrial internet of things: State-of-the-art and research challenges. *Sensors*, 19(21), p.4807.

Bhaiyat, H. and Sithungu, S. (2022, June) The Emergence of IIoT and its Cyber Security Issues in Critical Information Infrastructure. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1, pp. 46-51).

Caiza, G., Saeteros, M., Oñate, W. and Garcia, M.V. (2020) Fog computing at industrial level, architecture, latency, energy, and security: A review. *Heliyon*, 6(4), p.e03706.

Cao, K., Liu, Y., Meng, G. and Sun, Q. (2020) An overview on edge computing research. *IEEE access*, 8, pp.85714-85728.

Chalapathi, G.S.S., Chamola, V., Vaish, A. and Buyya, R. (2021) Industrial internet of things (iiot) applications of edge and fog computing: A review and future directions. *Fog/edge computing for security, privacy, and applications*, pp.293-325.

Gaidarski, I. and Minchev, Z. (2021) Insider Threats to IT Security of Critical Infrastructures. In *Digital Transformation, Cyber Security and Resilience of Modern Societies* (pp. 381-394). Springer, Cham.

Goswami, V., Jadav, P. and Soni, S.K. (2022) Review on How IIoT Has Revolutionized Greenhouse, Manufacturing and Medical Industries. In *Recent Advances in Mechanical Infrastructure* (pp. 179-192). Springer, Singapore.

Husain, B.H. and Askar, S. (2021) Survey on Edge Computing Security. *International Journal of Science and Business*, 5(3), pp.52-60.

Industrial Internet Consortium. (2019) *The Industrial Internet of Things Volume G1: Reference Architecture*. Available at: https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf [Accessed 19 November 2022]

Kebande, V.R., Karie, N.M. and Venter, H.S. (2018) Adding digital forensic readiness as a security component to the IoT domain.

Khan, W.Z., Rehman, M.H., Zangoti, H.M., Afzal, M.K., Armi, N. and Salah, K. (2020) Industrial internet of things: Recent advances, enabling technologies and open challenges. *Computers & Electrical Engineering*, 81, p.106522.

Koohang, A., Sargent, C.S., Nord, J.H. and Paliszkiewicz, J. (2022) Internet of Things (IoT): From awareness to continued use. *International Journal of Information Management*, 62, p.102442.

Lanz, Z. (2022) Cybersecurity Risk in US Critical Infrastructure: An Analysis of Publicly Available US Government Alerts and Advisories. *International Journal of Cybersecurity Intelligence & Cybercrime*, 5(1), pp.43-70.

Mirani, A.A., Velasco-Hernandez, G., Awasthi, A. and Walsh, J. (2022) Key Challenges and Emerging Technologies in Industrial IoT Architectures: A Review. *Sensors*, 22(15), p.5836.

Murray, G., Johnstone, M.N. and Valli, C. (2017) The convergence of IT and OT in critical infrastructure.

Patil, K., Gupta, S., Nair, A. and Gutte, V. (2020) Cloud, Fog and Edge Computing: Security and Privacy Concerns.

Parikh, S., Dave, D., Patel, R. and Doshi, N. (2019) Security and privacy issues in cloud, fog and edge computing. *Procedia Computer Science*, 160, pp.734-739.

Pivoto, D.G., de Almeida, L.F., da Rosa Righi, R., Rodrigues, J.J., Lugli, A.B. and Alberti, A.M. (2021). Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review. *Journal of Manufacturing Systems*, 58, pp.176-192.

Qiu, T., Chi, J., Zhou, X., Ning, Z., Atiquzzaman, M. and Wu, D.O. (2020). Edge computing in industrial internet of things: Architecture, advances and challenges. *IEEE Communications Surveys & Tutorials*, 22(4), pp.2462-2488.

Roman, R., Lopez, J. and Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, 78, pp.680-698.

Shalaginov, A., Iqbal, A. and Olegård, J. (2020), September. Iot digital forensics readiness in the edge: A roadmap for acquiring digital evidences from intelligent smart applications. In *International Conference on Edge Computing* (pp. 1-17). Springer, Cham.

Stoyanova, M., Nikoloudakis, Y., Panagiotakis, S., Pallis, E. and Markakis, E.K. (2020). A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues. *IEEE Communications Surveys & Tutorials*, 22(2), pp.1191-1221.

Sittón-Candanedo, I., Alonso, R.S., Corchado, J.M., Rodríguez-González, S. and Casado-Vara, R. (2019). A review of edge computing reference architectures and a new global edge proposal. *Future Generation Computer Systems*, 99, pp.278-294.

Xiao, Y., Jia, Y., Liu, C., Cheng, X., Yu, J. and Lv, W. (2019). Edge computing security: State of the art and challenges. *Proceedings of the IEEE*, 107(8), pp.1608-1631.

Zhao, Y., Yang, J., Bao, Y. and Song, H. (2021). Trustworthy authorization method for security in Industrial Internet of Things. *Ad Hoc Networks*, 121, p.102607.

Zhang, T., Li, Y. and Chen, C.P. (2021). Edge computing and its role in Industrial Internet: Methodologies, applications, and future directions. *Information Sciences*, 557, pp.34-65.

Cyber Threat Analysis in Smart Terminal Systems

Jussi Simola, Jouni Pöyhönen and Martti Lehto
University of Jyväskylä, Jyväskylä, Finland

Jussi.hm.simola@jyu.fi
jouni.a.poyhonen@jyu.fi
martti.j.lehto@jyu.fi

Abstract: Cyber threats create significant factors that challenge traditional threat prevention mechanisms in harbor areas and port terminals. It has been recognized that understanding security functionalities in the harbor area is based on a more traditional experience of what it requires. It is not enough that the maritime and harbor ecosystem repeats only physical security service routines regarding random checks of passengers and vehicles and customs functions on cargo and passenger transportation. Smart environments and infrastructures are widely expanded in urban areas and create more challenges if old practices are combined with new technologies and functionalities. Traditional threats have changed to a combination of threat types. While developing cyber or physical threats may evolve into hybrid threats, it may prevent everyday harbor activities so that damage can become long-lasting and harm business continuity management. Therefore, it is essential to analyze cyber threat factors in Smart Terminal Systems. The research provides cyber threat and vulnerability analysis and the main attack vectors in the Smart Terminal systems. This research belongs in Finland to the maritime Sea4Value (S4VF) research program that includes Smart Terminals (SMARTER) project for harbor's digitalization.

Keywords: Maritime Logistics, Smart Terminal Process, Cyber Security Analysis, Threats

1. Introduction

Global maritime and maritime logistic 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. The development of modern logistics depends entirely on a cyber environment that provides dynamic services. In Finland Smart Terminals (SMARTER) research project consists of port digitization by the end of 2023. 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 toward smart and autonomous maritime transportation. The project goals are conducted to the reduction of emissions by optimizing harbor operations and improving cargo and people flow while improving the experience for all stakeholders (DIMECC, 2020)

The structure of the project has three use cases. Those 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, including cyber security research actions in Work Package 4. (DIMECC, 2020)

The port cases digitalization means the development of Information and Communication Technology (ICT), Information Technology (IT), and Industrial Control System (ICS) or Operation Technologies (OT) solutions. Cyber threats create significant factors that challenge traditional threat prevention mechanisms in harbor areas and port terminals. In the future, maritime ports will become increasingly digitalized system systems (SoS). Cyber threats are system-level threats and will be needed to be coordinated like hybrid responses with other threats. Thus, it is necessary to address the comprehensive and relevant cyber threat analysis and security management aspects of the overall maritime solutions. It is crucial to enhance trustable services meaning the usability, reliability, and integrity of systems continuity within the operating environment. ENISA Threat Landscape Report 2016 emphasizes all elements covered within an attack on a business process. It means that not all artifacts/components used are IT-related; there are steps/procedures used within an attack, that are performed by having knowledge or information about the details of the business process at stake (ENISA, 2017).

This paper follows our previous research papers in this SMARTER study concerning cyber threat challenges in maritime logistics in in harbor areas and port terminals. The paper includes an analysis of the cyber threats in port systems, threat prevention requirements, and what prevention measures is needed to build up the comprehensive cyber security architecture for port services by exploiting the threat analyzes for cyber security measures for the SMARTER project. This study answers to the question, "Why cyber security management is so important in future smart ports system of systems environment?"

We have seen in a short timelapse how essential factor workable fairways and ports functions are. War in Ukraine and its derivative effects generated pressure on the vital functions of many countries and overall

continuity management. Economic balance becomes unstable easily when basic needs cannot be taken care of. Hybrid warfare consists, e.g., of exploited physical and cyber extortions. In Ukraine, we have seen that both elements have been used. Realized cyber threats create essential obstacles to maritime traffic and business continuity. In international business, we have to ensure that the functionalities of passenger and cargo traffic do not stop.

Criminal influences are trying to make our atmosphere unstable. Hybrid influencing is changed to hybrid warfare. As Simola & et al. (2021) wrote, fundamental risks effects decision-making culture and the whole cyber ecosystem. It is crucial to protect the harbor's cyber- and the physical environment against threats. Why it is important to try to understand attackers' motivation. Why is it important to create threat scenarios?

As terrorist actions in the Baltic Sea area and energy extortion are proven, every trade partnership between East and West is an opportunity and a risk. If decision-makers invest too much in the risk trade, we also create the possibility of risks spreading as a chain effect. One realizing threat affects another sector. A moment ago, leaking and exploding gas pipelines limited cargo vessels and passenger traffic in the Baltic Sea. After the terrorist attacks, Russia continued to spread disinformation (Financial Times, 2022).

2. Situational awareness in smart ports

As Endsley (1995) argues, the formation of situational awareness depends on the capabilities of perception and other individual factors. Individual thinking depends on many variable factors. Feelings affect our understanding and relationships related to the environment under consideration. An automated or semi-automated system requires capabilities to maintain situational awareness. It has been seen as necessary to monitor IT and OT systems itself and networks, but there is a need for harmonized situational awareness of Command and control -functionalities. Information exchange about cyber-physical threats has to be one of the system's core functions. SItuation center in the Port area needs a backup system if activities of the situation center are interrupted.

At present, a Cyber-Physical system is more often a system where IT and OT are linked or unified with Artificial Intelligence-based features. In this context, transportation, including trains, cars, trucks, airplanes, vessels, and boats, form the central functional wholeness where the main aim is to govern all necessary transportation-related issues effectively and securely. AI-based or combined solutions such as the Industrial Internet of Things (IIoT) create challenges for business continuity management. Figure 1 below illustrates systems of system-level approach interacts with other relevant factors (cyber and physical layers with human interaction).

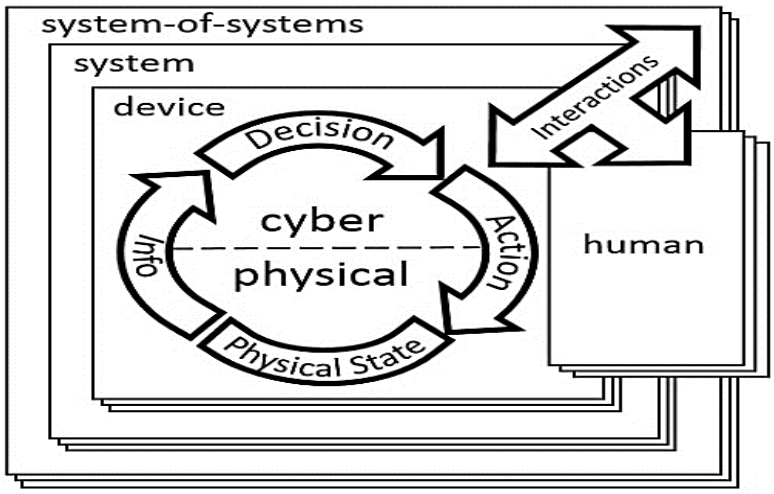


Figure 1: Cyber-Physical system of the system (NIST 2017)

Cyber-physical systems interact as part of the cyber ecosystem where cyber and physical stages combine work inputs. Human is still the primary decision-maker and actor that causes potential threats and challenges to business continuity management in many ways. It has been said that humans are the weakest players in work processes and procedures.

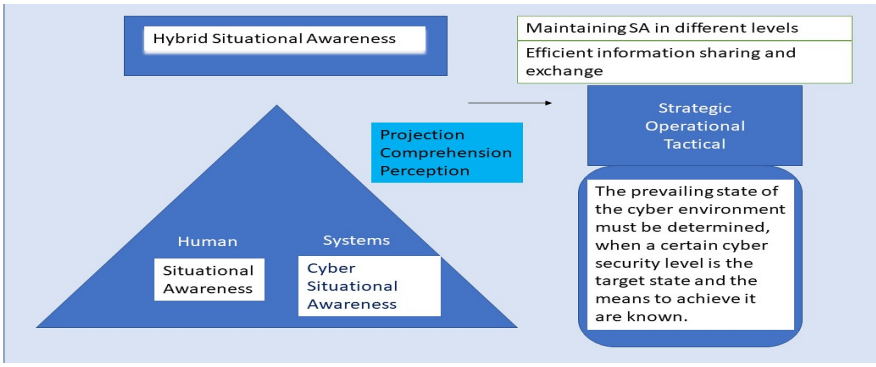


Figure 2: Formation of hybrid situational awareness in different levels

Stakeholders of the harbors have to create a preliminary risk assessment where every potential threat has been considered. As we have seen in previous studies related to Simola & et al. (2021) realized hybrid threats where cyber and physical risk elements are combined based on crucial human factors. It is not appropriate that risk classifications have been done separately from other risk assessments. Cyber risk assessment is one part of overall risk assessment. Figure 2 illustrates how the formation of hybrid situational awareness depends on determining the goal state of situational awareness in different analysis levels. Achieving and maintaining situational awareness at the goal state requires a coherent strategic, operational and tactical level of semantic functionalities where human and system-based information is shared understandably. The capability to understand threats and events creates a fundamental base to maintain everyday situational awareness. Prevention measures are equally important.

As Figure 3 illustrates, critical elements of the port have to analyze precisely. Port functionalities are an essential part of Critical Infrastructure. The risk management framework supports the decision-making process, in which critical infrastructure operators or partners undertake to cooperate in influencing the selection of risk management measures. It is designed to provide flexibility across all sectors, geographies, and across partners. It can be tailored to different operating environments and applies to all threats (DHS, 2013)



Figure 3: Critical Infrastructure Risk Management (DHS 2013)

The risk management concept allows operators of critical infrastructure (e.g., Security Operations Center) to focus on those threats and hazards that are likely to cause harm and to use approaches designed to prevent or mitigate the effects of these incidents. It also increases security and strengthens resilience by identifying and prioritizing actions to ensure the continuity of key operations and services and support enhanced response and recovery. In order to effectively risk management of critical infrastructure, port operators and stakeholders must identify (infra) the essential functions, systems, and networks necessary for their continued operation, taking into account the related dependencies and interdependencies. This dimension of the risk management process should also identify information and communication technologies that enable essential port services. Cybersecurity plans have to be a part of overall risk management activities, where policies, processes and procedures are defined and implemented.

We have many examples of how multi-activities in the maritime sector may expose stakeholders to cyber-physical influences, not only port activities. Harbors, ports, and fairways are crucial logistical parts of the critical infrastructure. NotPetya Malware affected, for example, A.P. Moller – Maersk In 2017. NotPetya was loaded onto one computer situated in the local office that was connected to the Maersk global network. Despite the fact that the company was not the intended target of the attack, it made all its applications and data unavailable. The whole logistical business chain and all operations were interrupted. The case proves that centralized ownership and control of functionalities create new business possibilities and also threats. If one employee would not open and respond to the infected email, or if the intrusion prevention mechanism had blocked the malware, the chain reaction would not have occurred (Stormshield, 2022).

3. Cyber-threat model for Smart Terminal

Identifying 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 and involve rapid technological developments, the struggle to meet them is ever-changing and increasingly complicated (Lehto, 2013).

According to Bodeau, McCollum & Fox, (2018), the word threat is used to refer to the adversary or the attack depending on the context. In the paper of "Cyber Threat Analysis in the Remote Pilotage System" (Kovanen, Pöyhönen, Lehto, 2021) is drafted a classification model for six threats based on motivational factors: cyber vandalism, cybercrime, cyber espionage, cyber terrorism, cyber sabotage, and cyber warfare. The motives can be reduced to their very essence: egoism, anarchy, money, destruction, paralysis, and power. The model was modified from Myriam Dunn Cavelty's structural model (Dunn Cavelty, 2010; Ashenden, 2011; Lehto, 2013).

There is a need to describe the actual attackers and their actions behind archetypes. Mitre's ATT&CK database (Strom et al., 2018) provides an updatable list of detected techniques, tools, and groups. The groups are tied to tactics, techniques, and tools they have been observed to use.

All the knowledge about the actor, attack techniques, and vulnerabilities must be available. It requires precise shared terminology and a method for transferring knowledge and information. The terminology should be able to address issues of varying abstraction levels and evolve with new concepts (Red Canary, 2021; ESET, 2019). ATT&CK also supports ICS and mobile system attack descriptions (Alexander, Belisle, & Steele, 2020; Strom et al., 2018).

Often cyber-attacks are associated, for example, with social or economic disputes, and the actions in the cyber domain may follow or be uncorrelated with events in the physical world (Gandhi et.al., 2011). Identifying the circumstances that might trigger an attacker archetype can be valuable in predicting heightened risk related to various situations.

Understanding the motivations and capabilities of different archetypes limits the number of scenarios and thus makes evaluation feasible for the defender. The motivation affects the attacker's targeting and methods. While a vandal seeks visibility by defacing a website, a spy wishes to stay unnoticed to gain information. The varying level of capability restricts some of the attackers from achieving their goals (Bodeau, McCollum & Fox, 2018).

The cyber vandalism may cause a chain reaction in a case where communication between VTS and an unknown vessel via a VHF channel or email is interrupted by a state actor. The arrival of an unknown vessel in a harbor area might trigger a chain reaction. The extended problem may arise if the same ICT systems are also in use in an enemy state's maritime transportation (Fintraffic, 2022). The awareness about the adversary that carried out the cyber-physical attack may be unclear. Cyber espionage can include business or political espionage. Political factors may arise from national or international issues. From the national side, hacktivism supporting strikes in the harbor could be one scenario. In the worst case, international tensions in the region could escalate to military cyber operations against vessel traffic. The parameters of the attacker archetypes are presented in Table 1.

Table 1: Attributes of the attacker archetypes. Capability is derived from Bodeau, McCollum and Fox (2018), and impacts are derived from Mitre (2020, 2022a, and Kovanen, Pöyhönen, Lehto, (2021).

	Vandalism	Crime	Espionage	Terrorism	Sabotage	Warfare operations
Motivation and goal	Political change based on personal political or ideological motives. Egoism gain	Making money through fraud or from the sale of valuable information. Financial gain	An economic, political, or military advantage. Information gain	Social instability and influencing political decision-making. Anarchy gain	Instability, chaos, political change, and infrastructure paralysis. Paralysis gain	A destructive attack on a nation's digital infrastructure. Political or military dominance

	Vandalism	Crime	Espionage	Terrorism	Sabotage	Warfare operations
Target	Digital services of governments and companies, individuals' information systems	Digital services of governments and companies, individuals' information systems	Data and information about governments and companies	Data and information about governments and companies. Critical infrastructure	Nation's critical infrastructure	Nation's critical infrastructure (civilian or military).
Capability	Acquired Attackers with moderate or limited expertise	Augmented Attackers with moderate or limited expertise	Advanced Attackers with very sophisticated or moderate expertise	Advanced Sophisticated attackers, capable of multiple, coordinated attacks	Integrated Very sophisticated attackers, capable of multiple, coordinated, continuous attacks	Integrated Very sophisticated attackers, capable of multiple, coordinated, continuous attacks
Trigger	A social event, an action of a company or an individual	The opportunity for economic gain	The need for political, economic, and military information	Cultural, nation's political or military actions	Testing own offensive cyber-attack capabilities, preparing hybrid or military operations	Achieving political or military objectives through military cyber operations

For Smart port, the model of cyber-threat folds around the threat archetypes and their features. The actions they make are described with a set vocabulary provided by ATT&CK.

4. Smart Terminal System-of-systems architecture

The case of port in the maritime transportation system includes processes to produce all needed services, as the ship approaches from an open sea via a fairway to berthing to a pier, general port services, port logistics, and connections to land transportation. There are identified the key elements associated with processes used in the smart terminal in figure 4. These are Activities, Stakeholders, Organizational relationships, Security dimensions, Security capabilities, and Criteria. It is also evident that this entity needs communication systems within process elements and electricity systems to support the functions of processes. In all cases of port processes, the information requirements and the amount of information needed are related to the reliability of safety and security services. Cyber security awareness and information should cover all process elements. Figure 4 presents typical port elements for cyber security investigation.

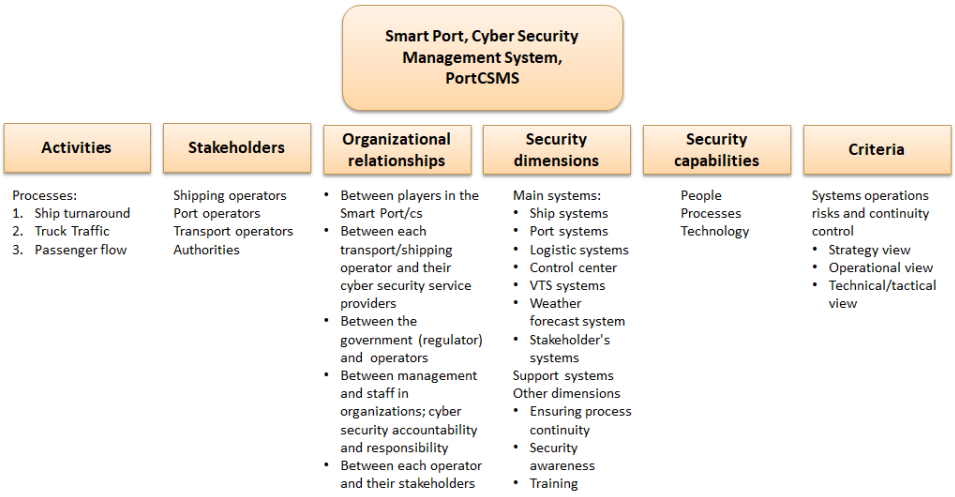


Figure 4: Elements of smart port cyber security management, Port CSMS.

In the SMARTER research project is acknowledged that there is a need for new technologies and solutions that are necessary to tackle the challenges set by the use cases (DIMECC, 2020). It is identified as a set of technologies and solutions that belongs to the terminology of Industry 4.0, like big data, data lake, data analytics, information fusion, AI, 5G, IoT, edge processing etc. It is reconditioned in the article of "Cybersecurity Risks and Automated Maritime Container Terminals in the Age of 4IR (Fourth Industrial Revolution)" by Peter Beaumont (2018) that the risks derived from the use of technology associated with the Fourth Industrial Revolution (4IR) are both real and dangerous unless appropriate control measures are implemented" (Beaumont, 2018). It is also important to recognize and have knowledge of legacy systems in use with a different timeframe of technology solutions. In that sense, the SoS environment of the port process is complicated from a technological point of view. In this SMARTER cyber security research, either new and legacy technologies should be covered in bout security dimensions, main and support systems of the case.

Smart ports cyber threats should need to be investigated and taking into account in system of systems (SoS) environment Information and Communication Technology (ICT) and Industrial Control Systems (ICS). The threat investigation of all systems should cover legacy and Industry 4.0 technology vulnerabilities because these systems are either critical for information security or cyber-physical effects. The findings from these systems provide the basement for evaluation and development of a cyber-threat analysis and are suitable for port cyber security architecture solutions.

5. Making threat analyses

We have used the Delphi method principle in order to make a relevant general threat analysis from the port systems. The members that have been involved in this analysis process are researchers and research methods. Cybersecurity experts from the research program advocate Delphi: “The Delphi method is an iterative process to increase consensus-building and at the end to have consensus among an experts from an examine case. The Delphi method is part of quantitative as a means to achieve an optimally reliable expert consensus.” It could have on one of three objectives (Garson, 2012): A) Forecasting future events, B) Achieving policy consensus on goals and objectives within organizations or groups, C) Identifying diversity in and obtaining feedback from stakeholders in some policy outcome.

Table 2 illustrates the results of Delphi method research on the port systems and subsystems. It has been done to forecast general or main future events conducted for the impacts and risk evaluation processes in the later phase of this research work. Cybersecurity researchers' and experts’ contributions are related to the main threats/attacks and the techniques. The evaluation has collected information about the threat agent involved, the attack that will be used, the vulnerability involved, and the possible approach of a successful exploit on the operation of the system. Categories of the systems are impossible to define clearly, because of the overlapping functions of the system's functionalities. Therefore same threat occurs in different system categories. It has also been challenging to determine clearly, was it IT or OT systems under the attack. Often cyber attacks affect both.

Table 2: Port systems and subsystems; main threats/attacks, related techniques (Mitre, 2020, 2022a, 2022b, 2022c, 2022d; Pöyhönen, J., Hummelholm, A., Lehto, M., 2022).

System/ Subsystems	Main threats/attacks	Software	ATT&CK technique	Threat prevention technologies
Ship systems	• Brute Force • Credential Theft • Physical Access • Denial of service (DoS) • Eavesdrop ping	• S0437 Kivars Remote services • Hardware or Software Keylogger	• T1110.001 -003 Password guessing, cracking, spraying, • T1111,T1621 MFA interception & request generation • T1556 Modify Authentication Process • T1552 Unsecured credentials • T0847 Replication Through Removable media	• M1032 Multi-factor Authentication • M1018 User Account Management • M1036 Account use policies • M1027 Password Policies • M1017 User Training • M1053 Data Backup • M0934 Limit Hardware Installation • M0928 Operating System Configuration
Port systems	• Man in the Middle • Jamming • Changing setpoints	• Hardware or Software Keylogger • S0504 Industroyer	• T1557.001-003.Adversary-in-the-Middle • T0860 Wireless compromise	• M1017 user training • M0931 Network Intrusion Prevention • M0947 Audit • M0802 Communication authenticity

System/ Subsystems	Main threats/attacks	Software	ATT&CK technique	Threat prevention technologies
	• Physical Access	• S0603 Stuxnet	• T1087 Account discovery	
Port logistic systems	• Physical Access • Virtual Access • Eavesdropping via malware	• Hardware or Software Keylogger • S0603 Stuxnet • S0604 Industroyer	• T0839 Module Firmware • T0842 Network Sniffing • T1485 Data Destruction • TO860 Wireless compromise	• M0808 Encrypt Network Traffic • M0806 Minimize wireless signals • M1053 Data Backup
Control center systems	• Credential Theft • Ransomware	• S0603 Stuxnet • S0125 Remsec • S0604 Industroyer • S1009 Triton	• T1110-003 Password guessing, cracking, spraying, • T1111,T1621 MFA interception & request generation • T1556 Modify Authentication Process • T1552 Unsecured credentials • T1562 Impair Defenses • T0803 Block Command Message	• M1031 Network Intrusion Prevention • M1018 User Account Management • M1022 Restrict File and Directory Permissions • M1024 Restrict Registry Permissions • M0804 Human user authentication • M0807 Network Allowlists • M1053 Data Backup
VTS system	• Attacks from Internet • Insider Attacks • DoS Attacks • API based Attack	• S0266 Trickbot • S0363 Empire	• T1059 Command and Scripting Interpreter • T1562 Impair Defenses • T1056.004 Input Capture: credential API hooking • T1071 Application layer • T1499 Endpoint DoS	• M1021 Restrict Web-based content • M1031 Network Intrusion Prevention • M1037 Filter Network Traffic
Weather forecast system	• Attacks from Internet • Insider Attacks • Credential Theft • DoS Attacks • DDoS • API based Attack • Ransomware Attacks	• S0125 Remsec • S0604 Industroyer	• T0918 Wireless sniffing • T1548 Abuse Elevation Control Mechanism • T1087 Account Discovery: Local Account • T0803 Block Command Message • T1486 Data Enrcytped for Impact	• M0948 Application Isolation and Sandboxing • M0950 Exploit Protection • M0930 Network Segmentation • M1047 Audit • M0814 Out-of Band Comm. Channel • Black Energy -malware toolkit • M1040 Behavior Prevention on Endpoint
Stakeholder`s systems	• Phishing • Brute Force	• S0367 Emotet • S0266 TrickBot •	• T1566.001 Phishing: Spearphishing Attachment • T1087 Email discovery	• M9031 Network Intrusion prevention • M0930 Network segmentation • M1036 Account use policies
Power supply systems	• DoS - Impacts on the ICS (combined effects)	• S0604 Industroyer • S0603 Stuxnet	• T0855 Unauthorized Command Message • T0816 Device Restart	• M0801 Access Man. • M0802 Communication authenticity • M0937 Filter Network Traffic • M0930 Segmentation • M0813 Software Process and Device Authentication

In an ICT environment, it is essential to collect a knowledge base about the general-level port activities and how the systems are connected. For that analysis, Mitre has created a usable framework in this context (Mitre, 2022d). In the cyber-physical environment, we have to notice also signal-based threats such as Electromagnetic interference (EMI), which is an increasing threat factor in the wireless communication environment.

Because of digitalization, combined IT and OT systems often mean that the operational technology is linked to the system connected to the public network. Crucial threats from external factor cause challenges in port facilities because OT and IT Systems often communicate with each other in internal and external networks. It is essential how employees use their internet access. Adversaries may deploy vulnerabilities by scanning and sniffing the target network, e.g., by using spear-phishing attachments in emails or packet sniffer or wireless analyzer so that crucial port processes (for example, cargo handling or access control system) may stop. The malware can also get to the internal network via a USB plug-in. Human errors with poor identity & access management are essential threats. Therefore accurate responsibilities and tasks for key crew members have to be done. If port operations are suspended for a long time, it will affect the entire logistics supply chain and expenses will increase.

The similarity of the threats indicates the importance of protecting information and its sharing and storing practices. For that aim, the CIA technique consists of guides for policies to form information security within an organization. Confidentiality, Integrity, and Availability (CIA) feature. Confidentiality means that information is only accessible to those involved. Integrity or correctness of information means that the information must be true and correct. Availability means that information is available when you want to use the data subject's data.

Elements of the CIA should be implemented throughout the cybersecurity-managed IT services and other security-related management. The right to privacy or the rights of the data subject required by data protection cannot be fulfilled without the implementation of the data security attributes. Cyber threats create significant challenges for trusted information sharing. Protecting port activities require common guidelines for cyber security rules.

For example, ISO/IEC 29134:2017 gives guidelines for a process on privacy impact assessments (PIA) and a structure and content of a PIA report. It is applicable to all types and sizes of organizations, including public companies, private companies, government entities, and not-for-profit organizations (ISO, 2017).

The Mitre framework uses three terms in an ICS environment; *denial*, *loss*, and *manipulation*. Denial is a condition that occurs only while the attack is active. Loss refers to the sustained loss of an asset that continues after the active malicious interaction has ceased. Manipulation alters the asset and can be either loud and easy to detect or subtle and longer sustained. According to previous research (Kovanen, Pöyhönen & Lehto, 2021), we have described the attacks as follows:

- **Manipulation of view** is a more inconspicuous attack type than denial or loss of view. Slightly falsified data are harder to detect than missing data. Consequently, the affected system operator loses correct situational awareness. The effect spreads to all connected systems and operators using the manipulated view.
- **Denial of Service attacks** can be carried out by affecting the endpoint or the network that leads there. In either case, the service is unavailable for use.
- **Data destruction, data encrypted for impact, disk wipe and service stop,** prevent the use of the data and services. System shutdown/reboot can be used to make systems inaccessible faster by, for example, rebooting after wiping the master boot record.
- **Loss of Safety** is dangerous, especially with cyber physic systems as the result may cause injuries or death when the safety mechanism of a system is disabled. Even a threat of this type of circumstance can delay reaction to other impact types if a human operator is not able to initiate countermeasures due to a fear of unsafe working conditions.
- **Data manipulation** is harder to detect than data destruction if the manipulation is subtle. Systems and operators can continue to act but they base their decisions on false data.

6. Conclusion

The Finnish smart port research program concerns port services for ship turnaround, truck traffic, and passenger processes. Smart ports are enormous mixtures of legacy technology and the development of new apps. The potential for cyber attacks on systems is always present and evolving along with digitalization. All emerging vulnerabilities associated with the interconnection of ICT and ICS/OT layers in smart terminals must be thoroughly assessed.

The threat investigation of ICT and ICS/OT systems of smart ports poses a severe impact on maritime processes in harbor areas. The analysis indicates that system-level thinking is necessary when the purpose is to understand the dependencies of the threat sources and attack vectors. One threat factor affects another that may combine different threat levels from each other. Distributed systems with separate departments or functions may not provide protection in this case. Outsourced services create their challenge. Therefore, it is essential to deploy standards and other industry rules. Without a common understanding of procedures and processes regarding workflow, threat prevention is not possible to take control. Information security and technology have to be based on common standards, certificates, and protocols. Recommendations that we have on shipboard networks should be of the same level in the cyber port ecosystem as well. Business continuity management has to take into account every technical solution that is planned to deploy in this maritime ecosystem.

Port owners form an essential entity for management when the cyber ecosystem has to manage understandably and easily. There is a challenging example where control of the port entity is divided abroad and added challenges arise from a subsidiary arrangement. The system of system-level design depends on how the ecosystem is governed. Coordination and control of the activities influence how ICT and ICS/OT technologies support each other in the port area and other connected functions. Official National Cyber Security Center Traficom has a Havaro mechanism that monitors internet traffic by using a commercial Security Operation Center that gathers different companies under the sector-based communities. It is recommended to get all port actors involved and connected to the umbrella of the service. Additionally, it is necessary to increase attention to educating employees on cybersecurity activities. Human resource management affects how stable or weak human resources are in a digitalized cyber-physical environment.

This paper provides a research approach to threat identification of the system element of the smart terminal processes. The research approach uses the system of systems (SoS) thinking. The findings of the study are related to the coverage of multiple main threats in port services. Information flows in and between systems as well as electricity supplies in the port process.

References

Alexander, O., Belisle, M., and Steele, J. (2020). “Mitre ATT&CK® for Industrial Control Systems: Design and Philosophy”.

Ashenden, D. (2011). “Cyber Security: Time for Engagement and Debate”, Proceedings of the 10th European Conference on Information Warfare and Security, the Institute of Cybernetics at the Tallinn University of Technology.

Bodeau, D.J. and McCollum, C.D. (2018). “System-of-systems Threat Model”, The Homeland Security Systems Engineering and Development Institute (HSSEDI) MITRE, Bedford.

Bodeau, D.J., McCollum, C.D. and Fox, D.B. (2018) “Cyber Threat Modeling: Survey, Assessment, and Representative Framework”, Mitre Corp, Mclean.

DHS, (2013). NIPP 2013 - Partnering for Critical Infrastructure Security and Resilience.

DIMECC Oy, 2020. DIMECC Sea4Value/Smart Terminals (SMARTER). Project proposal for One Sea – autonomous maritime ecosystem.

Dunn Cavelty, M. (2010). “The Reality and Future of Cyberwar”, Parliamentary Brief, 30 March 2010, [online], www.parliamentarybrief.com/2010/03/the-reality-and-future-of-cyberwar.

Endsley, M.R. (1995). A taxonomy of situation awareness errors, human factors in aviation operations. Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP). 3. 287-292.

ENISA. (2017). Threat Landscape Report 2016, 15 Top Cyber-Threats and Trends.

ESET. (2019). “The MITRE ATT&CK Framework: Everything You Need to Know in Under 60 Minutes ”, [online], https://www.eset.com/us/business/resources/webinars/the-mitre-attck-framework-everything-you-need-to-know-in-under-60-minutes-1/.

Financial Times. (2022). Sabotage of gas pipelines a wake-up call for Europe, officials warn. https://www.ft.com/content/ad885fea-035f-4b93-98e7-c75da2c308f8

Fintraffic. (2020). GOFREP. https://www.fintraffic.fi/sites/default/files/2021-09/GOFREP_MG_2021_09_03.pdf

Gandhi, R., Sharma, A., Mahoney, W., Sousan, W., Zhu, Q. and Laplante, P. (2011) “Dimensions of Cyber-Attacks: Cultural, Social, Economic, and Political”, IEEE Technology and Society Magazine, Vol. 30, No. 1, pp 28–38. doi: 10.1109/MTS.2011.940293

Garson, G. D. (2012). The Delphi method in quantitative research. Asheboro, NC: Statistical Associates Publishers. Available from: https://faculty.chass.ncsu.edu/garson/PA765/delphi.htm, retrieved 25.1.2022

International Organization for Standardization (ISO). (2017). ISO/IEC 29134:2017 guidelines for privacy impact assessment. Retrieved from https://www.iso.org/standard/62289.html

Kovanen, T., Pöyhönen, J. & Lehto, M. (2021a). Cyber Threat Analysis in the Remote Pilotage System. Presented in ECCWS 2021 - 20th European Conference on Cyber Warfare and Security. 24th - 25th June 2021, Chester, UK. doi: 10.347190/EWS.21.067

Lehto, M. (2013). "The Cyberspace Threats and Cyber Security Objectives in the Cyber Security Strategies", International Journal of Cyber Warfare and Terrorism, Vol. 3, No. 3, pp 1–18.
Mitre. (2020). "ATT&CK®for Industrial Control Systems",[online], https://collaborate.mitre.org/attackics/index.php/Impact
Mitre. (2022a). "Impact", [online], https://attack.mitre.org/tactics/enterpise
Mitre. (2022b). "Software" [online], https://attack.mitre.org/software.
Mitre. (2022c). "ATT&CK Matrix for Enterprise", [online], https://attack.mitre.org/.
Mitre. (2022d). "Attack Navigator", [online], https://mitre-attack.github.io/attack-navigator/
NIST. (2017). Special Publication 1500-201 Framework for Cyber-Physical Systems: Volume 1, Overview
Pöyhönen J., Hummelholm A., Lehto M. (2022). Cybersecurity risk assessment subjects in information flows. 21st European Conference on Cyber Warfare and Security, 16th – 17th June 2022, Chester, UK, pages 222-230. doi: 10.34190/eccws.21.1.263
Red Canary. (2021). "2020 Threat Detection Report", [online], https://redcanary.com/threat-detection-report/.
Simola J., Lehto M., Rajamäki J. (2021). "Emergency Response Model as a part of the Smart Society", Proceeding of the 20th European Conference on Cyber Warfare and Security, pp. 382-391. doi: 10.34190/ews.21.079
Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G. and Thomas, C.B. (2018) "Mitre Att&ck: Design and Philosophy", Technical report.
Stormshield. (2022) Port cyberattack: hackers & maritime cybersecurity, https://www.stormshield.com/news/cybermaretique-a-short-history-of-cyberattacks-against-ports/

Towards a Scientific Definition of Cyber Resilience

Sidney C. Smith
DEVCOM Army Research Laboratory, Adelphi, MD, USA

sidney.c.smith24.civ@army.mil

Abstract: Cyber resilience must be improved. Improving cyber resilience requires the quantitatively measuring it. However, before cyber resilience can be measured, it must first be scientifically defined. An effort to discover a consensus among researchers as to the scientific definition of resilience, in general, and cyber resilience, specifically, revealed that no such consensus exists. Experts from several disciplines agree that the word resilience is becoming a meaningless buzz word. This paper reviews the literature to establish the current state of the scientific definition of resilience. It briefly surveys the literature to discover what makes a valid scientific definition. It reviews and analyses the historic scientific use of resilience to discover the path from its original meaning to its current diverse and conflicting meanings. These concepts are decomposed using a genus-differentia analysis untangling the various connotations and separating the related but different concepts. Based upon this analysis, a proposal is made that resilience is part of a family of properties under the umbrella of tenacity. This family includes resistance, resilience, persistence, and perseverance. Finally, an initial operational definition of cyber resilience based upon key performance parameters under stress is proposed.

Keywords: Cyber Resilience, Scientific Definition, Stress, Strain, Resistance, Persistence

1. Introduction

Many are engaged in various efforts to improve cyber resilience. To improve cyber resilience, it must be measured (Kott and Linkov, 2021). "A valid definition of a concept is a prerequisite to valid measurement (Locke, 2003, p. 416)." Thus the formulation of a valid definition of cyber resilience is necessary before it may be quantitatively measured. Several authors have observed that progress in many areas of research is being hindered by a lack of definitions (Locke, 2003; Wacker 2004; Podsakoff, MacKenzie, and Podsakoff, 2016). Locke (2003) observed that some researchers fail to define the concept that they are studying, and others search the literature for a consensus on a definition without ever evaluating that definition. Hibberd observed that some concepts suffer from having too many conflicting definitions and cites three papers to substantiate his claim (Mulligan and Scherer, 2012; Maul, 2017; Rossiter, 2017, cited in Hibberd, 2019, p. 30). Brtis (2021, p. 1) observed, "One expert claims that well over 100 unique definitions of resilience have appeared." Aburn, Gott, and Hoare (2016, p. 1) found, "there is no universal definition of resilience." The concept of resilience is suffering from too many conflicting definitions.

1.1 Challenge

Locke (2003) asserts that good definitions are the epistemological foundation of scientific progress and sloppy definitions are a major factor retarding intellectual progress. Hibberd (2019, p. 29) observes that without rigorous definitions, "no discipline advances from vagueness and ambiguities of a less than technical language and the follies that result." Björck et al. (2015, p. 1) observed, "In order for cyber resilience to gain momentum also as an academic research subject, it is important to define the term." Hoffman and Hancock (2017, p. 564) observed that there has been considerable interest in resilience and that, "concepts that come to the forefront of concern in this manner are often diluted, simply to become the next 'flavor of the month' through both overselling and uncritical use." Further, "In such evolutions or revolutions, the definition of terms often proves to be a problematic issue that frequently threatens to derail important conceptual progress" (Hoffman and Hancock, 2017, p. 565). Hosseini, Barker, and Ramierez-Marquez (2016, p. 49) conducted a review of definitions and measures of resilience and concluded, "The review of resilience definitions indicates that there is no unique insight about how to define the resilience." Gigerenzer (2017) points out that building upon the work of others requires good scientific definitions. He further stated, "The practice of using the same label for logically and operationally different phenomena impedes progress" (Gigerenzer, 2017, p. 137).

1.2 Frustration

A literature review search for a consensus of the definition of resilience, in general, and cyber resilience, specifically uncovered the only consensus on the definition of resilience is that there is no consensus on the definition of resilience. Klien, Nicholls, and Thomalla (2003) conducted a literature review of the definition of resilience. They observed that in the time from 1973 until 2003, resilience has transformed from a "straightforward concept used only in mechanics" to a "complex multinterpretable concept with contested definitions and relevance" (Klien et al., 2003, p. 40). Dillon (1947, p. 207) observed, "There seems to be almost

as many different definitions of `resilience' as there are authors." Norris et al. (2008), conducted a literature review of the definition of resilience. They observed that the definitions in use in the social and psychological sciences had diverged so far from the original definition that they wondered if perhaps they should have created their own term (Norris et al., 2008). Reid and Botterill (2013, p. 31) conducted an overview of the multiple meanings of resilience in the literature. The key conclusion from their research is, "the term is highly ambiguous, it is used for different purposes in different contexts and in some cases the understandings of the term are diametrically opposed (Reid and Botterill, 2013, p. 31)." Arghandeh et al. (2016, p. 1), searched the literature for a definition of cyber-physical resilience in power systems and concluded, "There is no clear and universally accepted definition of cyber-physical resilience for power systems." Hosseini, Barker, and Ramierez-Marquez (2016) conducted a review of definitions of resilience in systems engineering. They discovered "Many overlap with a number of already existing concepts such as robustness, fault-tolerance, flexibility, survivability, and agility, among others (Hosseini et al., 2016, p. 48)." Xue, Wang, and Yang (2018) explored the science of resilience with a critical review and bibliometric analysis. They summarized the definition of resilience in social and ecological, engineering and disaster, and economic and organizational behavior domains and concluded, "The definition of resilience is still an important research area (Xue et al., 2018, p. 503)." Cottam et al. (2019) completed a structured literature review of the definition of resilience in engineered systems. After focusing only on resilience in engineering systems and excluding definitions from psychology, ecology, and socio-ecological system, they considered 54 papers. They "found no standard definition of resilience and the definitions found included a variety of means in their definitions of resilience (Cottam et al., 2019, p. 26)." Many researchers have thoroughly reviewed the definition of resilience in the literature and concluded that over the years the definition has expanded to the point of meaningless jargon (Xue et al., 2018).

1.3 Contribution

Having established the frustration in the literature about the definition of resilience in general and cyber resilience specifically, the contribution of this paper is to briefly review the literature for what constitutes a valid scientific definition. The history of the definition of resilience is then examined. This knowledge is then applied in a genus differentia analysis to propose a valid scientific definition of cyber resilience that will lend itself to a quantitative measurement. Although many have systematically reviewed the literature for the definition of resilience, I know of no other effort to apply the rules for definitions or conduct a genus differentia analysis of resilience.

1.4 Organization

Section 2 will review the literature searching for guidelines for scientific definitions. Section 3 will discuss the approach. Section 4 will trace the meaning of resilience from the Latin roots through the various fields of scientific endeavor, attempt to untangle the various conflicting definitions of resilience into the separate but related components using a genus differentia analysis, and consider an operational definition of cyber resilience. Section 5 will provide a conclusion and discussion of future work.

2. Background

Since the ancient Greeks, scientists and philosophers have struggled to find valid definitions of critical concepts. They principally used the genius-differentia method of definition assigning a concept to a family and describing what makes it different from other members of that family (Bayer, 1998). Locke (2003, p. 416) observes, "A valid definition of a word (i.e., concept) accomplishes two things: (a) it ties the concept to reality, and (b) it distinguishes the concept from other concepts." Robinson (1950) talked about real definitions as analysis, synthesis, and improvement of concepts. When he used the word analysis, he meant breaking a thing down into its component parts (Robinson, 1950). When he used the word synthesis, he was talking about discovering that the thing in question is part of a larger whole (Robinson, 1950). Real definitions are used to refine or improve concepts by substituting a similar concept which is superior (Robinson, 1950). Wacker (2004) provides eight rules for formal conceptual definitions. He encourages clarity and simplicity. Definitions should not use vague or ambiguous terms and they should also not expand the concept or make it less exclusive. Further new hypotheses should not be introduced in the definition (i.e., the definition should not include instances where only 'good' events happen (Wacker, 2004). Hibberd (2019) lists seven things that a scientific definition is not. She emphasizes that a scientific definition is concept or real definitions by stating that it is not one of several nominal definitions. Like Locke she asserts that a scientific definition must tie a concept to reality. She also concludes that scientific definitions are not operational definitions (Hibberd, 2019). An operational definition defines a

concept by describing the technique used to measure it (Bridgman, 1927). Podsakoff MacKenzie, and Podsakoff (2016) provide ten recommendations for better concept definitions. They stress clarity and completeness like Wacker. They also warn against the use of various types of nominal definitions like Hibberd (Podsakoff et al., 2016). Cottam et al. (2019) observes that definitions should not contain the means of obtaining the object. They relate this to the "what not how" principle (Date, 2000). This illustrates the principle expressed by General Patton, "Never tell people how to do things. Tell them what to do, and they will surprise you with their ingenuity (Patton and Harkins, 1947)."

3. Approach

Many have conducted a literature review of resilience looking for a consensus and finding none. Rather than conducting yet another literature review, this article traces the history of the meaning of resilience in an effort track its evolution and understand the expansion that has led to the current state of meaninglessness. The standard account of the scientific definition has used the genus-differentia model that has been employed since the ancient Greeks. Although there is no longer a requirement that scientific definitions be stated in this form, this form of definition satisfies Locke's (2003) requirements. This exercise allows the examination of the words in common use surrounding the concept of resilience and establishes its place in that context. Then consideration is given to the operational definition of cyber resilience. Although a scientific definition is more than an operational definition, it is appropriate to give some thought as to how to measure this concept.

4. Results

4.1 History

4.1.1 Latin Root

The word resilience comes to English from the French *resiler* and the Latin *resilire*. It is composed of the *re* meaning back and *salire* meaning jump (Brown, 1993). The word carries the meaning of jumping back, recoiling, springing back, or resuming an original position (Brown, 1993; Hosseini, Barker, and Ramierez-Marquez, 2016).

4.1.2 1858 Civil Engineering and Mechanics

In 1858 Rankine used the word to describe the effects of stress and strain on solids (Rankine, 1858). In his discussion, he defined *stress* as the force applied to an object, and *strain* as the alteration observed in the object (Rankine, 1858). *Elasticity* is when the object recovers its original shape when the stress is withdrawn. *Plasticity* is when the object does not retain its original shape when the stress is withdrawn (Rankine, 1858). *Fracture* is the point at which the object is divided into parts. *Toughness* is the greatest strain which the body will bear without fracture (Rankine, 1858). *Resilience* is the quantity of mechanical work required to produce the proof strain (Rankine, 1858). Later this concept was expressed in a stress–strain graph, see figure 1. In this context, the properties of stiffness, resilience, and toughness have neither a negative nor a positive connotation.

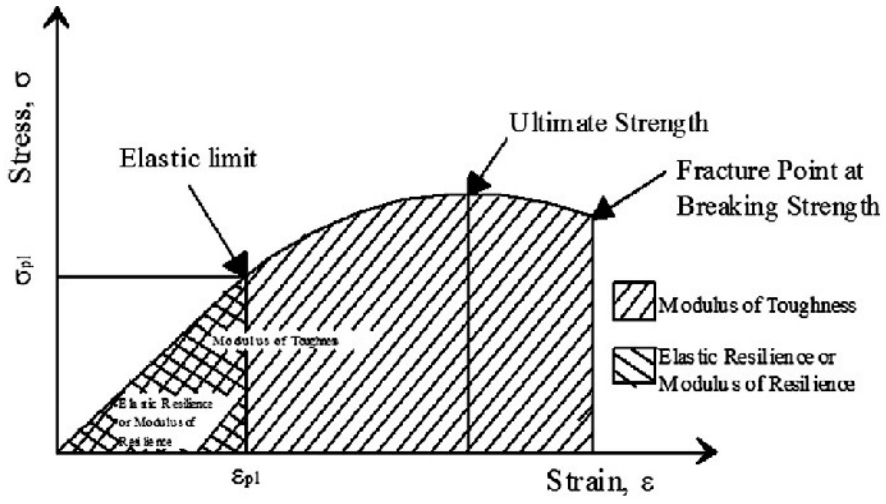


Figure 1. Stress–strain graph (Ahmmad et al., 2014).

4.1.3 1947 Fibres and Fabrics

In 1947 when Dillon wrote his article, "Resilience of Fibers and Fabrics," the definition of resilience in this context had grown and diversified until, "there seem to be almost as many definitions of `resilience' as there are authors. (Dillon, 1947, p. 207)" The materials that Rankine (1958) studied obeyed Hooke's law; however, some fibers and fabrics do not. Hooke's law states that within the elastic range the stress and strain scale linearly. This made the measure of resilience as a function of work problematic. Even though the basic distinction between stiffness, resilience, and toughness exists, Dillon (1947, p. 212) finishes his work saying, "The obvious conclusion from this study is that resilience is a much abused and poorly defined term and that much remains to be learned about its significance and the factors controlling it, in single fibers, bulked fibers, and fabrics."

Hoffman (1948, p. 141) defines resiliency as, "as the capability of a substance to return to its original state at some later time after the removal of the deforming stress." He adds time to stress and strain as a component of resilience. The materials that Rankine considered immediately returned to their original state once the stress was removed. Wool more gradually returns to its original state. His proposed definition is, "a stress–strain-time property of a material, characterizing the completeness of recovery from deformation and varying in kind with the modulus of elasticity and rate of recovery (Hoffman 1948, p. 148)."

Resilience gains a positive connotation because fabrics with too much stiffness are uncomfortable to wear, and fabrics without resilience stretch becoming baggy in the joints. The word resilience begins to mean fit for a particular purpose (e.g., "An experimental tire made of woolen cords overheated badly; therefore, wool is not resilient (Hoffman, 1948, p. 141).")

4.1.4 1973 Ecology

In 1973 Holling applied the term resilience to ecological systems. He stated, "Resilience determines the persistence of relationships within a system and is a measure of the ability of these systems to absorb changes of state variables, driving variables, and parameters, and still persist (Holling, 1973, p. 17)." Holling faced problems beyond the scope of previous research into resilience. How does one measure the ability of a system to return to a previous state when the state of the system is constantly changing? Ecologies may be said to exist in a stable equilibrium. Resilience measures the amount of stress that the system can absorb before a different stable equilibrium is established. Xue, Wang, and Yang (2018, p. 500) point to Holling's work as the beginning of an explosion in resilience research stating, "The definition of resilience in different categories are all derived from the ecology domain."

The distinction between stiffness and resilience starts to blur. Gruemm (1976) proposed a measurement of resilience based upon stability theory using differential topology where the state space of the system is divided into basins with each basin containing an attractor. A system is considered resilient against a certain stress if the system returns to the same basin (Gruemm, 1976). This treatment distinguishes fracture or extinction, plastic change when the system moves to a different basin, and elastic change when the system returns in the original basin. It does not distinguish no change from elastic change.

4.1.5 1971 Psychology

The study of resilience in psychology begin in orthopsychiatry. While studying individuals with high-risk factors for developing psychosis, it was discovered that some of the subjects thrived despite very high-risk factors. These individuals were originally called invulnerable, but this was later replaced with the term resilient (Garmezy, 1971; Egeland, Carlson, and Sroufe, 1993; Masten, 2001; Glantz and Slobada, 2002).

The line between resisting and rebounding is further blurred, and the connotation of growth through or despite trauma is introduced. The positive connotation of the word resilience grows significantly. Resilience is now seen as a very desirable quality.

4.1.6 1988 Social

Wildavsky (1988) considered the resilience of social systems and observed that stress may be necessary for the development of resilience. Adger (2000) merged the concepts of social and ecological resilience creating the concept of social-ecological systems (Berkes, Colding, and Folke, 2008). Efforts are undertaken to measure and improve the resilience of communities to disasters (Bruneau et al., 2003; Rose, 2007). Olsson et al. (2025) explain why the concept of resilience as positive growth after stress gains traction in this field and why resilience as rebound to a previous state is unappealing.

Natural and man-made disasters also illustrated the need for resilient organizations, and the means of building resilience began to creep into the definition of resilience. For example, the National Research Council (2012, p. 1) defined resilience as, “the ability to prepare and plan for, absorb, recover from and more successfully adapt to adverse events.”

4.1.7 2009 Cyber

These trends continued when resilience was introduced to the cyber domain. The scope of resilience expanded to encompass “the ability to provide and maintain an acceptable level of service in the face of faults and challenges to normal operations (Goldman et al., 2011, p. 236).” The means to achieving cyber resilience creep into the definition, “Cyber resilience (or resiliency) is the ability to anticipate, withstand, recover from, and adapt to adverse conditions, stresses, attacks, or compromises on cyber resources (Graubart and Bodeau, 2016, p. 1; Bodeau et al., 2018, p. iii).”

4.1.8 Summary

Since the early 1970s, the definition of resilience has expanded from elastic deformation to include every aspect of strength (i.e., stiffness, spring, and toughness.) It has grown beyond the original concept of strength to include positive change or growth. Phillips and Chao (2022, p. 7) go so far as to say, “Resilience is not simple survival, bounce-back, or homeostasis.” The implication here is that resilience no longer means what it has always meant. Most of Wacker’s (2004, p. 637) rules for good formal definitions deal with clarity; however, the sixth rule states, “Definitions should not make any term broader. New definitions should not expand the concept to make it broader and less exclusive.”

As the connotation of the word resilience gains in desirability and many are striving to encourage the improvement of resilience, means like adaptability and planning have been added to the definition. However, Cottam et al. (2019) observed that definitions should not contain the means of obtaining the object.

4.2 Genus-differentia

At a high level one can say that resilience is a quality attribute. Miller (1996) defines quality as, “fitness for use” with the understanding that this has two aspects: “product features that meet customer needs and freedom from defects (Miller, 1996, p. 3 citing Juran 2003).” Barbacci et al. (1995, p. 3) define software quality, “Software quality is the degree to which something possesses a desired combination of attributes (e.g., reliability, interoperability).” However, the concept of quality has expanded to cover almost every aspect of a system. Barbacci et al. (1995) decompose quality into performance, dependability, security, and safety.

Since Laprie (2008) defines resilience in terms of dependability, “dependability when facing changes,'' dependability seems to be appropriate aspect of quality to pursue. Dependability is defined as the set of systems properties “that allows us to rely on a system functioning as required (Littlewood and Strigini, 2000).” This definition is further refined, “Dependability is that property of a computer system such that reliance can justifiably be placed on the service it delivers. Dependability has several attributes, including availability, reliability, safety, confidentiality, integrity, maintainability (Barbacci et al., p. 15)''

It is important to differentiate between dependability and security since both dependability and security are concerned with availability, confidentiality, and integrity. When Bishop (2003) defines computer security, he talks about requirements, policies, and mechanisms. Security is about having the mechanisms to implement the policies that fulfil the requirements aligning well with the “fitness for use” aspect of quality. Dependability is about these mechanisms working correctly and continuing to work correctly during adverse conditions aligning well with the “free from defects” aspect of quality.

Woods (2015) describes four concepts of resilience: rebound, robustness, graceful extensibility, and sustained adaptability. Robustness is “the ability of a system to function correctly in the presence of disturbances (Rungger and Tabuada, 2016, p. 2108).” A robust system is one where small disturbances lead to small deviations and the effect of sporadic disturbances disappear over time (Tabuada et al., 2014). A robust system would be more resilient than a fragile system; however, a system could be robust and still ineffective in a hostile environment since the only requirement is that it be able to recover from sporadic disturbances.

The concepts of graceful extensibility and sustained adaptability map directly to the dependability concept of maintainability. These are certainly desirable qualities and could contribute to the resilience of a system if measured over a system’s lifetime; however, they would contribute very little if resilience is measured over a

mission. Therefore, they would be better considered as a means of implementing resilience than as part of the definition of resilience itself. This leaves the concept of rebound, which maps directly to the lexical definition of resilience.

Barbacci et al. (1995, p. 16) defines reliability as "a measure of the ability of a system to keep operating over time." Since reliability applies without any unforeseen or external stress, a more precise word for reliability in the face of stress might be strength or tenacity. Tenacity may be further decomposed into stiffness or resistance, resilience, and toughness or persistence. Resistance speaks to a system's ability to withstand stress without a perceptible degradation in performance. Resilience speaks to a system's ability to absorb stress elastically, that is performance will perceptibly degrade but return when the stress is removed or nullified. Toughness or persistence speaks to a system's ability to absorb stress plastically, that is the system will continue to function at an acceptable level of performance in the face of stress. It also carries the idea of bending but not breaking easily. In addition, there is the special case of perseverance or adaptability. Perseverance is plastic change; however, this change improves the tenacity of the system. An example would be recovering from an illness where the recovery builds immunity providing enhance resistance in the future. All of these are important concepts that many of the definitions of resilience in the literature confuse making it difficult to distinguish between them.

Resilience then is a property of tenacity. Where tenacity describes the ability of a system to withstand stress without failure. Resilience specifically is the elastic response to stress where the system is measurably impacted but returns to the former state when the stress is removed or nullified

4.3 Operational Definition

If we restrict the definition of cyber resilience to the cyber stress that a system can absorb through elastic deformation, the most straightforward measure would be cyber stress. Although there are some discreet measures of cyber stress, there is no continuous measure that would be necessary for a quantitative measure of cyber resilience. Although the stress itself cannot be quantitatively measured, the strain or degradation of key performance parameters (KPPs) can be. The family of cyber strength may be measured by the difference in KPPs without and with cyber stress.

5. Conclusion

The recent explosion in the popularity and varying definitions of resilience have left it a panchreston. If the concept of resilience is to have any scientific significance, its meaning must be better focused. The proposed method of focusing the meaning of resilience is to understand it as part of a family of properties that describe the relationship between stress and strain on a system. Cyber resilience is the ability of a cyber system to recover from stress that causes a reduction of performance. This may be measured by the degradation of KPPs caused by the stress. There is still much work to be done to take this general concept of cyber strength in general and cyber resilience particularly and mathematically model and quantitatively measure it.

Acknowledgements

The author would like to thank the Office of the Under Secretary of Defense for Research and Engineering for funding the Quantitative Measurement of Cyber Resilience project, and Dr. Alexander Kott for his preliminary work and continuing advice and direction.

References

Aburn, G., Gott, M. and Hoare, K. (2016) "What is resilience? An integrative review of the empirical literature", *Journal of Advanced Nursing,* Vol. 72 No. 5, pp. 980-1000.

Adger, W.N. (2000) "Social and ecological resilience: are they related?", *Progress in Human Geography,* Vol. 24, No. 3, pp. 347-364.

Ahmmad, R., et al. (2014) "Ductility performance of lightweight concrete element containing massive palm shell clinker", *Construction and Building Materials,* Vol. 63, pp. 234-241.

Arghandeh, R., et al. (2016) "On the definition of cyber-physical resilience in power systems", *Renewable and Sustainable Energy Reviews,* Vol. 58, pp. 1060-1069.

Barbacci, M., et al. (1995) "*Quality attributes",* Software Engineering Institute, Carnegie Mellon University, Pittsburg, PA.

Bayer, G. (1998) "Classification and explanation in Aristotle's theory of definition", *Journal of the History of Philosophy,* Vol. 36, No. 4, pp. 487-505.

Berkes, F., Colding, J. and Folke, C. (2008) "*Navigating social-ecological systems: building resilience for complexity and change",* Cambridge University Press.
Bishop, M. (2003) "What is computer security?" *IEEE Security & Privacy,* Vol. 1, No. 1, pp. 67-69.
Björck, F., et al. (2015) Cyber resilience--fundamentals for a definition. In: *New contributions in information systems and technologies,* Springer, pp. 311-316.
Bodeau, D.J., et al. (2018) *Cyber resiliency metrics catalog,* MITRE Corp, Bedford, MA.
Bridgman, P.W. (1927) *The logic of modern physics.* 3 ed. Macmillan, New York, NY.
Brown, L. editor. (1993) Resilience, In: *The new shorter oxford english dictionary on historical principles,* Oxford University Press, Oxford.
Brtis, J., Jackson, S., and Cureton, K. (2021), *Systems resilience,* Systems Engineering Body of Knowledge (SEBok), https://www.sebokwiki.org/w/index.php?title=System_Resilience&oldid=62047.
Bruneau, M., et al. (2003), "A framework to quantitatively assess and enhance the seismic resilience of communities", *Earthquake Spectra,* Vol. 19, No. 4, pp. 733-752.
Cottam, B., et al. (2019) *Defining resilience for engineered systems,* Canadian Center of Science and Education.
Date, C.J. (2000) *What not how the business rules approach to application development,* Addison-Wesley, Upper Saddle River, NJ.
Dillon, J.H. (1947). "Resilience of fibers and fabrics", *Textile Research Journal,* Vol. 17, No. 4, pp. 207-213.
Egeland, B., Carlson, E. and Sroufe, L. A. (1993) "Resilience as process", *Development and Psychopathology,* Vol. 5, No. 4, pp. 517-528.
Garmezy, N. (1971) "Vulnerability research and the issue of primary prevention. *American Journal of Orthopsychiatry",* Vol. 41, No. 1, p. 101.
Gigerenzer, G. (2017), "A theory integration program", *Decision,* Vol. 4, No. 3, pp. 133-145.
Glantz, M. and Slobada, Z. (2002) Analysis and reconceptualization of resilience, In: *Resilience and Development,* Springer, pp. 109-126.
Goldman, H., McQuaid, R. and Picciotto, J. (2011) *Cyber resilience for mission assurance,* IEEE, Waltham, MA.
Graubart, R. and Bodeau, D. (2016) *Cyber resilience metrics: key observations,* MITRE CORP, McLean, VA.
Gruemm, H.R. (1976) *Definitions of resilience,* Definitions of resilience, Laxenburg, AT.
Hibberd, F.J. (2019) "What is scientific definition?", *Journal of Mind & Behavior,* Vol. 40, No. 1, pp. 29-52.
Hoffman, R.M. (1948) "A generalized concept of resilience", *Textile Research Journal,* Vol. 18, No. 3, pp. 141-148.
Hoffman, R.R., and Hancock, P.A. (2017) "Measuring resilience", *Human Factors,* Vol. 59, No. 4, pp. 564-581.
Holling, C. S. (1973) "Resilience and stability of ecological systems", *Annual Review of Ecology and Systematics,* Vol. 4, No. 1, pp. 1-23.
Hosseini, S., Barker, K., and Ramierez-Marquez, J.E. (2016) "A review of definitions and measures of system resilience", *Reliability Engineering & System Safety,* Vol. 145, pp. 47-61.
Juran J.M. (2003) *Juran on leadership for quality,* Simon and Schuster.
Klien, R.J., Nicholls, R J. and Thomalla, F. (2003) "Resilience to natural hazards: how useful is this concept?", *Global Environmental Change Part B: Environmental Hazards,* Vol. 5, No. 1, pp. 35-45.
Kott, A. and Linkov, I. (2021). To improve cyber resilience, measure it, *arXiv preprint arXiv:2102.09455.*
Laprie, J.-C. (2008) *From dependability to resilience,* IEEE, Anchorage, AK.
Littlewood, B. and Strigini, L. (2000). *Software reliability and dependability: a roadmap,* ACM, LImerick, Ireland.
Locke, E.A. (2003). Good definitions: the epistemological foundation of scientific progress, In: *Organizational behavior: the state of the science,* Lawrence Erlbaum Associates, Mahwah, NJ, pp. 395-425.
Masten, A.S. (2001) "Ordinary magic: resilience processes in development", *American Psychologist,* Vol. 56 No. 3, p. 227.
Maul, A. (2017) "Rethinking traditional methods of survey validation", *Measurement: Interdisciplinary Research and Perspectives,* Vol. 15, No. 2, pp. 51-69.
Miller, W.J. (1996) "A working definition for total quality management (TQM) researchers", *Journal of Quality Management,* Vol. 1, No. 2, pp. 149-159.
Mulligan, K. and Scherer, K.R. (2012) "Toward a working definition of emotion", *Emotion Review,* Vol. 4, No. 4, pp. 345-357.
National Research Council, (2012) *Disaster resilience: a national imperative,* National Academies Press, Washington, DC.
Norris, F.H., et al. (2008) "Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness", *American Journal of Community Psychology,* Vol. 41, No. 1, pp. 127-150.
Olsson, L., et al. (2015) "Why resilience is unappealing to social science: theoretical and empirical investigations of the scientific use of resilience", *Science Advances,* Vol. 1, No. 4, pp. 1-11.
Patton, G.S. and Harkins, P.D. (1947). *War as I knew it,* Houghton Mifflin, Boston, MA.
Phillips, F.Y. and Chao, A. (2022) "Rethinking resilience: definition, context, and measure", *IEEE Transactions on Engineering Management,* pp. 1-8.
Podsakoff, P.M., MacKenzie, S.B. and Podsakoff, N.P. (2016) "Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences", *Organizational Research Methods,* Vol. 19 No. 2, pp. 159-203.
Rankine, W.J.M. (1858) *A manual of applied mechanics,* Richard Griffin and Company, London.
Reid, R. and Botterill, L.C. (2013) "The multiple meanings of `resilience': an overview of the literature", *Australian Journal of Public Administration,* Vol. 72, No. 1, pp. 31-40.
Robinson, R. (1950) *Definition,* Oxford University Press, Amen House, London.

Rose, A. (2007) "Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions", *Environmental Hazards,* Vol. 7, No. 4, pp. 383-398.

Rossiter, J.R. (2017) "Optimal standard measures: comment on Matthews et al.(2016)", *The American Psychologist,* Vol. 72, No. 5, pp. 489-490.

Rungger, M. and Tabuada, P. (2016) "A notion of robustness for cyber-physical systems", *IEEE Transactions on Automatic Control,* Vol. 61, No. 8, pp. 2108-2123.

Saxe, J.G. (1899) *The poetical works of John Godfrey Saxe,* Houghton, Mifflin and Company, Boston, MA.

Spence, J. (1820) *Observations, anecdotes, and characters of books and men,* John Murray, London, UK.

Tabuada, P., et al. (2014) "Towards robustness for cyber-physical systems", *IEEE Transactions on Automatic Control,* Vol. 59, No. 12, pp. 3151-3163.

Wacker, J.G. (2004) "A theory of formal conceptual definitions: developing theory-building measurement instruments", *Journal of Operations Management,* Vol. 22, No, 6, pp. 629-650.

Wildavasky, A.B. (1988) *Searching for safety,* Transaction Publishers, Piscataway, NJ.

Woods, D.D. (2015) "Four concepts for resilience and the implications for the future of resilience engineering", *Reliability Engineering & System Safety,* Vol. 141, pp. 5-9.

Xue, X., Wang, L. and Yang, R.J. (2018) "Exploring the science of resilience: critical review and bibliometric analysis", *Natural Hazards,* Vol. 90, No. 1, pp. 477-510.

Implications of Cyberbiosecurity in Advanced Agriculture

Simone Stephen[1-2],Keitavius Alexander[1-2], Lucas Potter[3], Xavier-Lewis Palmer[2-3]
[1]Old Dominion University, Norfolk, USA
[2]CySecSol, Franklin, USA
[3]BiosView, Oswego, USA

sstep008@odu.edu
kalex019@odu.edu
lpott005@odu.edu
xpalm001@odu.edu

Abstract: The world is currently undergoing a rapid digital transformation sometimes referred to as the fourth industrial revolution. During this transformation, it is increasingly clear that many scientific fields are not prepared for this change. One specific area is agriculture. As the sector which creates global food supply, this critical infrastructure requires detailed assessment and research via newly developed technologies (Millett et al, 2019; Peccoud et al, 2018) . Despite its fundamental significance to modern civilization, many aspects of industrial agriculture have not yet adapted to the digital world. This is evident in the many vulnerabilities currently present within agricultural systems, as well as the lacking and fragmented nature of policy dictating cybersecurity stances– the field which intersects both cybersecurity and biosecurity to protect several areas within life sciences (Murch et al, 2018; Duncan et al, 2019; U.S. Department of Agriculture, 2022) . These looming oversights create dangers to advanced agricultural systems, which in turn poses risk to businesses, economies, and individuals. While there are various methods to reduce these risk factors, they ultimately depend on the careful consideration of cyberbiosecurity (CBS) by all involved. This includes the system developers, equipment engineers, and especially the end users - all of us. A conscientious team-effort can work to diminish risks and ultimately provide a safer environment for advanced agriculture and all who depend on it. This analysis explores numerous vulnerabilities within the system of advanced agriculture, discusses potential solutions to the escalating risks they present, and considers the achievable future of an advanced agricultural system which further implements the role of CBS.

1. Introduction

The United States is one of the world's leading food producers, responsible for more than $176 billion of agricultural exports in 2020 (M. Shahbandeh, 2022). The trend within the consistently high production of US agriculture is that of centralization. While the number of farms has declined from 6.8 million in 1935 to 2 million today, the average farm size has gone from 1.5 to 4.5 hundred acres (USDA 2022). This has been enabled by the changing technology of the time, the modern implications of which will be the major topic of this paper.

The cyber revolution is rapidly shaping the world and the agricultural field has begun to implement several new technologies including the internet into its environment. While these new developments have become increasingly essential, the integration of these technologies have also become a source of possible danger. With the emergence of multiple threats in the form of malicious attacks, unwanted intrusions, and sabotage within cyberspace, cybersecurity has become of critical importance within the agricultural field. Recognition of this and many other threatened fields of science has prompted the development of CBS. This new discipline is commonly described as a convergence of biosecurity, cybersecurity, and cyber-physical security. The field encompasses advancing and implementing the protection of life and medical sciences, cyber, cyber-physical, logistics network and infrastructure systems (Murch et al., 2018). As a critical infrastructure, the agricultural field has demonstrated a growing necessity for this discipline. Applications consist of safeguarding intellectual property, data integrity, availability, operation of essential systems, and overall quality of end products (Murch et al, 2018). Other functions include the restriction of access to valuable resources and data, remaining in accordance with standards, regulations, and other important documents, as well as the prevention of financial or reputational damages (Carneiro et al). These aspects could all have significant consequences not only for the industries and organizations themselves, but also their many consumers. Attacks on systems within the agricultural field could disturb the food supply chains already made fragile by the unique challenges presented during the COVID-19 pandemic (Rejeb et al., 2020). This could ultimately result in significant damage to the bioeconomy and communities (Drape et al, 2021). These factors suggest that there is a significant need for cybersecurity.

This also applies to the area of advanced agriculture. For the purposes of this work, Advanced Agriculture is defined as the technology-enabled development of data-driven agricultural systems involving food production, technological food processes, animal feed, bioenergy, agrofuels, and bio-based products (Kanski, 2020). This is often innovative research with the aim of maximizing efficiency, productivity, availability, and sustainability of farming methods (Kanski, 2020). However, despite advanced agricultural systems also having a significant effect on global food supply chains and the bioeconomy, they are remarkably ill-equipped for cyber-attacks. This is evident in an article by Christine Whitt in 2021 stating that at least 80% of the agricultural and food operation in the United States is generated by small family farms, ranches, and food producers (Whitt, 2021). This is an issue because these agricultural sectors receive practically no consideration regarding the security of cyber/IT infrastructure (Whitt, 2021, cited in Drape & Murch, 2022). This is especially an issue as advanced agriculture also commonly involves information sharing for the purpose of facilitating improved data interpretation. Restricting this exchange of data complicates the design of produce generated for specific consumer demands as well as the value chain, and can ultimately heighten CBS risks (Duncan et al., 2020). CBS risks in this field are numerous and increasing every day as technological developments continue to be implemented. These vulnerabilities could allow for significant damage to the confidentiality, integrity, and availability of digital information within agricultural systems. It is crucial for CBS to be further implemented into advanced agricultural systems to safeguard the bioeconomy, food supply chain, and ultimately communities around the world. This work will examine the vulnerabilities within these systems, possible threats, the implications of these threats, and discuss solutions towards reducing risks and aspects of future integration of CBS within advanced agriculture. The major areas of these issues in the farming environment are noted in figure 1 below.

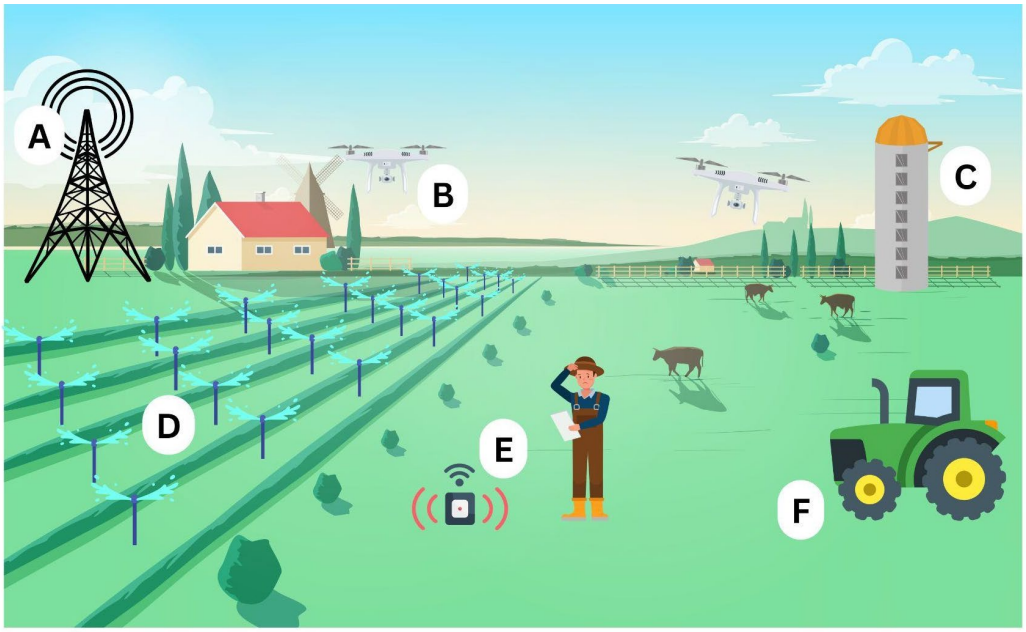


Figure 1: This is an image of a modern farm implementing advanced agricultural aspects. (A) Cyber attacks on the electrical power grids could cease all farm operations that require electricity. (B) Cyber attacks targeting drones used on farms. These attacks could result in the deterioration of the quality of the soil, increasing the difficulty of producing and maintaining a large quantity of agricultural goods (Croptracker). (C) Cyber attacks could target the agricultural sensors and climate controls of storage silos. (D) Cyber attacks on water distribution/irrigation systems. (E) Cyber attackers could target agricultural sensors tracking information regarding: the amount of sunlight received, amount of humidity, compression and density of soil, climate of environment surrounding, soil air penetration, farming activity conducted on soil, and similar data regarding the soil (Geopard, 2022). (F) Cyber attacks targeting technology used as tools or equipment on farms could result in farming becoming more difficult.

2. Hardware

With digital transformation affecting more fields than before, the agricultural field continues to develop technological mechanisms and devices to incorporate into their systems. This includes a multitude of sensor nodes used to attain environmental and livestock data, unmanned vehicles and other robotics used for precise automation of routine field tasks, and the integration of IoT platforms for the use of controlling, interacting, and monitoring of a system of connected devices (Bacco et al., 2019). These developments improve many aspects of the system such as efficiency, productivity, and manageability. For this reason, these new mechanisms and equipment have largely become a necessity. However, the implementation of these new devices also creates many weaknesses within the system. The first component of vulnerability within these systems to be discussed is hardware. In advanced agriculture at the field level, this can take the form of actuators, sensors, and vehicles. These can be used for monitoring critical environmental parameters, executing farming functions, and delivering automated transport (Alahmadi et al., 2022; Saiz-Rubio & Rovira-Más, 2020). While these technologies provide great utility and convenience, not enough attention is given to these devices. Many of these hardware devices do not include an integrated security mechanism to examine their usage, status, or memory access which is essential in securely storing cryptographic keys (Alahmadi et al.,2022). This in turn makes them vulnerable to certain threats.

One of the most common known threats are trojans which are a kind of malicious misuse that can impact system hardware components through covert modification of devices. This could severely alter the intended function of devices or even go as far to stop a function or resource from operating in a denial-of-service attack (Jang-Jaccard & Nepal, 2014). A possible example of this threat causing problems in an agricultural context could be stuxnet-like malware causing the overwatering of crops within a farming system (Wolf, 2014). Another increasing exploitation of hardware is the illegal cloning of devices to contain malware which can cause harm and information leaks (Jang-Jaccard & Nepal, 2014). A possible instance of this would be an attacker discreetly administering a cloned fertilizer spreader with irregular distribution intended to cause under or over fertilization of crops. Another common threat to hardware being side-channel attacks. These are attacks made through exploiting information extracted from the data leakage of a physical cryptosystem. This can take the form of exploiting timing, power consumption, electromagnetic emissions, and thermal signatures which deliver valuable data capable of revealing the encryption standard as well as the keys to crack the encryption (Alahmadi et al., 2022). This can cause the leakage of very sensitive data which can be used or manipulated by attackers. Ultimately, these many vulnerabilities within the hardware of advanced agriculture can cause serious harm and demand to be addressed and appropriately secured.

3. Software

Despite hardware attacks having greater capacity for system influence, the manipulation of software vulnerabilities constitute a significant amount of cyber-attacks. These commonly occur via malware, a form of malicious software that can cause inconvenience, loss, or damage. It is usually loaded into a system, unknown to the system owner, to compromise it for an attacker's gain. It can come in the form of worms, viruses, ransomware, spyware, and bot executables that can be injected to affect nodes and devices (Jaccard & Nepal, 2014; Leclair & Ramsay, 2015). Other causes of software vulnerability within agricultural systems are attacks on applications, such as software update attacks which are executed by interrupting the update procedure of installed software. Another example is buffer overflow attacks, which take advantage of coding errors to overwrite memory and compromise systems. Application attacks also include SQL injection, which is the use of code injection techniques to deceive a database server into executing malicious SQL code that is injected into a database input field. A possible example of this would be an attacker using SQL injection to penetrate and exploit a logistics system into transporting crops to different places than intended or to steal private data from storage systems. These attacks all have the massive potential to cause damage to computer systems (Yazdinejad et al., 2021).

As farms and related research facilities continue to integrate technology, the risk of these attacks increases significantly. Yet, despite recognition of these dangers, many agricultural systems are not prepared for protecting against these attacks. It has been found that in some farming systems, anti-virus software was only installed on most computer systems instead of all, while in others, there were different anti-virus programs downloaded on different workstations. It was also found that some of these systems utilized the same computer systems for both on-farm integration and web-browsing. This also takes place within the agricultural research field in which many projects promote the improved use of technology, but neglect safety in their models

(Windows, 2019). This is evident in how a research study by Kim et al. (2018, cited in Windows, 2019) had been written on the use of farm as a service, sensors, AI usage, and the Internet of Things without any mention of the possible threats and dangers presented. It has also been seen in a paper by Walter et al. (2017, cited in Windows, 2019) that even if dangers are incorporated into articles, they are seen as incidental instead of potential attack points to be taken advantage of by attackers . This demonstrates that despite the rising integration of technology within these agricultural systems, there is a blatant lack of security practices set in place. Ultimately, these vulnerabilities within both the hardware and software applications of advanced agricultural systems demand the implementation of safety measures to create a more secure environment.

4. Threats/Exploitation

4.1 Threats to confidentiality, integrity, availability

The agriculture industry is reliant on secure technologies to maintain, produce, and transport livestock, farm goods, and more. In the future farmer-free farms may exist and be operated by artificial intelligence and robotic systems with only minimal human intervention. Taking insights from Demestichas et al (2020), smart farming setups can have numerous points at which privacy losses and data breaches can occur Demestichas et al., 2020). For instance, a smaller farmer may rely on a mobile computing-based platform for predictions on what crops to grow and when, and false information may cause real-world consequences insofar as crop failure or suboptimal yields. However, a larger agricultural operation may be a more completely autonomous operation, like a smart farming environment. In either case, the real-world necessity and purpose of smart operations make suspensions of network activity dire (Demestichas et al., 2020). Operations with more points of failure and less layering of protective operations can have more opportunities for hardware-based attacks and malfunctions (Demestichas et al., 2020). Digital adversaries seek to gain unauthorized access to machinery and software to gain access to information, databases, and intellectual property. As this industry develops and adapts and advances more capable and hi-tech, developments in cyber defenses should be established. The Frontiers article, “Assessing the Role of Cyberbiosecurity in Agriculture, A Case Study”, stated that:“Smart technologies, while beneficial, have the potential to be exploited by hackers to disrupt the farms using them and the downstream users relying on the supply chain. Potential risks that are attributed with precision agriculture and smart technologies include: false sensor data, data and machinery access control, and data encryption (i.e., ransomware attacks)” (Chi et al., 2017, cited in Drape, 2021).

Cyber threats may also attempt to disrupt agricultural production through means of advanced cyber attacks such as artificial intelligence attacks. AI attacks will search for vulnerabilities more efficiently, and quickly than a cyber attack led by a human. AI cyber attacks also develop massive databases on vulnerable corporate smart farms. A new generation of cybercriminals has emerged, one that is both subtle and secretive, which will influence the future of cybersecurity. “The new generation of cyber threats will be smarter and capable of acting independently with the help of AI. Future cyberattack methods will be able to be aware of their surroundings and make informed decisions based on the target environment” (Guembe et al., 2022). By learning the pattern of repeated behavior of agriculturally focused attacks and their targets, defense actors can make more educated decisions on how to counter malicious ones who may aim to infiltrate, block access, and change or delete information utilized to compute and control operations related to agricultural operations.

4.2 Who benefits from these Exploitations

Exploitation of security vulnerabilities of softwares, machinery, and digital agriculture tools can be beneficial to numerous parties who hold a stake in the bioeconomy, from local to national (brands, governments, cyber criminal organizations, and industries) who view categories and subcategories of agriculture as a threat to their own. Resources such as vegetation, seeds, irrigation systems, cattle in livestock, foods, and fuels are valuables to threats. Stolen data from vulnerable digital and hardware infrastructure and framework can be utilized to attack individuals, companies, corporations, and more. Insights from Alahmadi et al (2022) suggest that relationships between producers of agricultural products and those who trade them are important security intersections to examine. Through the monitoring of parameters that trading competitors use to gauge production efficiency, an actor can isolate weak links in a region’s bioeconomy (Alahmadi et al., 2022). Competitors use data gained through cyber attack to exploit vulnerable farms more and develop larger scale attacks to farms with more advanced cyber defenses. On a national scale, this can be a route to disrupting competing bioeconomies in order to change the state of competition. Expanded to a war-time scenario, this

could be a route by which countries asymmetrically compete with their enemies, to the dismay and harm of citizens caught in between. Competition in this vein, through attacking bioeconomies and their supports, have the potential to shorten conflicts, but a huge problem with this is dealing with the after-effects restoring the damaged systems. It is hoped that sufficiently designed agricultural security systems can mitigate this possibility.

5. What Implications Could Mean For Development/Safety of Advanced Agriculture

An agricultural industry that does not address current and future threats is likely to face harsh consequences from adversaries that have prepared these potential threats. The exploitation of automation and remote controls of machinery that can be used to maintain and supervise farms could be used to sabotage produce and ruin a potential energy source in the future. The agricultural industry could become a key player in the energy industry. Photosynthesis plays an important role in agriculture not only for the production of energy rich foods; also for future avenues to combat climate change. Synthetic and genetically engineered photosynthesis could be a prominent source of oxygen and CO2. Researchers are eyeing new plant engineering techniques that could improve crop yields and improve food security. (Baslam et al., 2020.) However, with growing concerns of climate and renewable energy, energy and developmental resources located and operated at smart farms could be hacked. Technologies including genetically engineered vegetation could face threats who seek to steal and ruin production though hardware or software attacks. Synthetic photosynthesis and genetically modified plants could be the future of agriculture; however, development and resource allocation must take place to secure placement for new technologies in agriculture. This risk must also be balanced in the midst of macro-level induced issues that endanger agricultural resources. For example, during the COVID-19 pandemic simple vegetation seedlings were difficult to come across. This has caused seed shortages which are expected to also affect the value chain (Nchanji et al., 2021). Similar environmental conditions could occur in a future experiencing extreme consequences of climate change. During these incidents, cyber threats can exploit technological vulnerabilities to gain access and control resources such as the genetically modified seedlings. Advanced genetic engineering technologies could be used as a weapon to steal resources from farms. One hypothetical, future example may be in the form of an engineered DNA stealing parasite that could be a threat to the confidentiality of targeted farms. In the case of Claude dePamphilis, professor of biology, parasitic plants who stole and weaponized genetic information from host plants were found (Swayne, 2016). What may follow from this possibility is the ability of malicious actors one day enlisting biological agents to steal or compromise unique GMO intellectual property. This hypothetical type of attack would not be one that is easily detectable via legacy visual means such as cameras or satellites. A solution to this may lie within passive monitoring systems installed throughout farms. This could also be supported by active monitoring methods through sequencing of samples taken through swabbing, plant cuttings, insect collection, and routine soil examinations that monitor changes in genetic material within these farm environments. We may look to innovations shown in dynamic, rapid integrated monitoring systems in the future.

6. Developing Thoughts on the Future of Advanced Agriculture

As the fourth industrial revolution proceeds to shape the world, the advanced agricultural field will continue to implement technology. To provide the world's growing population with sufficient food and energy products while also combating climate change, the advanced agricultural sector must continue to integrate technology to maximize efficiency. The World Population Prospects project (2015, cited in Goedde et al., 2020) that the population is to reach 9.7 billion by 2050. This in turn, prompts the need for a 70% increase in calories available for consumers according to a report by the United Nations on World Resources Institute (2013, cited in Goedde et al., 2020). The United Nations has also stated that the water supply will be 40% below global water demand by 2030 (2016, cited in Goedde et al., 2020). These projections in combination with the rise of environmental pressures and urge for sustainable agricultural practices suggest the need for further technological integration to resolve these imminent issues. This can take the form of further implementation of smart devices, plant propagation and cultivation technologies, intelligent automated mechanisms, and effective data analysis (Dhar, 2022). These technologies can also consist of more advanced tools like sentinel satellites and drones to provide photos for mapping of spatial variability. These maps could then be used to measure things like crop harvest, humidity levels, topography, and resources in the area (Gorli, 2017). These innovations can aid in ultimately increasing development, diminishing waste, and improving profit margin within the advanced agricultural field (Eashwar & Chawla, 2021). However, they may just be the beginning of an increased utilization of technology. It has been predicted by OnFarm, the creator of a connected farm IoT platform, that the average farm in 2050 will generate about 4.1 million data points per day. This is a 190,000-point increase from the year 2014 (Gorli, 2017).

This demonstrates that the amount of technology implemented into agricultural systems will increase dramatically in years to come. However, the question remains as to whether the advanced agricultural field will be able to properly integrate CBS practices in preparation for this impending innovation. As cyber-attacks are inevitable, CBS will be essential to the future success and security of the agricultural supply chain and bioeconomy, and it would be advantageous to increase and improve security to avert successful attacks rather than in reaction to them (Van der burg et al., 2019).

7. Conclusion

The recent developments in agricultural systems and other similar life sciences have established new technologies which in turn introduce a multitude of new cybersecurity threats. Recognition and detection of these susceptibilities fall under CBS. This work first goes on to describe the relevance, significance, and many vulnerabilities within advanced agricultural systems. Threats to confidentiality, availability, and integrity and potential actors of these threats were then presented to provide a landscape of the many dangers within these spaces. Beyond this, the possible consequences of these threats affecting further development of advanced agricultural systems was established to demonstrate the significance of applying relevant CBS perspectives. The work then finally discussed upcoming world developments prompting the demand for further innovation within agriculture, providing additional incentive for CBS practices in the future. Ultimately, this paper highlights the dangers involved in unprotected advanced agricultural systems and calls for further analysis and awareness of CBS while strongly suggesting the need for solutions.

References

Alahmadi, Adel N., et al. “Cyber-Security Threats and Side-Channel Attacks for Digital Agriculture.” *Sensors*, vol. 22, no. 9, 2022, p. 3520., https://doi.org/10.3390/s22093520.

Ahmad W, Rasool A, Javed AR, Baker T, Jalil Z. Cyber Security in IoT-Based Cloud Computing: A Comprehensive Survey. Electronics. 2022; 11(1):16. https://doi.org/10.3390/electronics11010016

Bacco, Manlio, et al. “The Digitisation of Agriculture: A Survey of Research Activities on Smart Farming.” *Array*, vol. 3-4, 2019, p. 100009., https://doi.org/10.1016/j.array.2019.100009.

Baslam, M., Mitsui, T., Hodges, M., Priesack, E., Herritt, M., Aranjelo, I., & Saenz-Saez, A. (2020). Photosynthesis in a Changing Global Climate: Scaling Up and Scaling Down in Crops. Frontiers in Plant Science, 11(1664-462X). https://doi.org/10.3389/fpls.2020.00882

Carneiro, Renata, et al. “Cyber Attacks in Agriculture: Protecting Your Farm and Small Business with Cyberbiosecurity.” *VCE Publications | Virginia Tech*, 14 July 2021, https://www.pubs.ext.vt.edu/content/pubs_ext_vt_edu/en/FST/FST-387/FST-387.html.

Chi, H., Welch, S., Vasserman, E. and Kalaimannan, E., 2017. A framework of cybersecurity approaches in precision agriculture. In proceedings of the ICMLG2017 5th International Conference on Management Leadership and Governance (pp. 90-95). Reading, UK: Acad. Conf. Publ. Int.

Croptracker (n.d.). Croptracker - Drone Technology In Agriculture. [online] www.croptracker.com. Available at: https://www.croptracker.com/blog/drone-technology-in-agriculture.html.

Demestichas, K., Peppes, N., & Alexakis, T. (2020). Survey on Security Threats in Agricultural IoT and Smart Farming. Sensors, 20(22), 6458. https://doi.org/10.3390/s20226458

Dhar, Payal. “Cybersecurity Report: ‘Smart Farms’ Are Hackable Farms.” *IEEE Spectrum*, IEEE Spectrum, 19 Sept. 2022, https://spectrum.ieee.org/cybersecurity-report-how-smart-farming-can-be-hacked.

Drape, Tiffany, et al. “Assessing the Role of Cyberbiosecurity in Agriculture: A Case Study.” *Frontiers in Bioengineering and Biotechnology*, vol. 9, 2021, https://doi.org/10.3389/fbioe.2021.737927.

Drape, Tiffany A., and Randall Murch. “Leveraging Cyberbiosecurity to Safeguard Agriculture and Food.” *VTechWorks Home*, Virginia Tech, 1 Oct. 2022, https://vtechworks.lib.vt.edu/handle/10919/112168.

Duncan, S. E., Reinhard, R., Williams, R. C., Ramsey, F., Thomason, W., Lee, K., ... & Murch, R. (2019). Cyberbiosecurity: A new perspective on protecting US food and agricultural system. *Frontiers in bioengineering and biotechnology*, *7*, 63.

Duncan, Susan E., et al. “Securing Data in Life Sciences—a Plant Food (Edamame) Systems Case Study.” *Frontiers in Sustainability*, vol. 1, 2020, https://doi.org/10.3389/frsus.2020.600394.

Eashwar, S, and Paras Chawla. “Evolution of Agritech Business 4.0 – Architecture and Future Research Directions.” *IOP Conference Series: Earth and Environmental Science*, vol. 775, no. 1, 2021, p. 012011., https://doi.org/10.1088/1755-1315/775/1/012011.

GeoPard Agriculture. (2022). “What are precision agriculture sensors.” [online] Available at: https://geopard.tech/blog/what-are-the-types-of-sensors-used-in-agriculture/#.

Goedde, Lutz, et al. “Agriculture's Connected Future: How Technology Can Yield New Growth.” *McKinsey & Company*, McKinsey & Company, 11 Nov. 2020, https://www.mckinsey.com/industries/agriculture/our-insights/agricultures-connected-future-how-technology-can-yield-new-growth.

Gorli, Ravi, and G. Yamini. "Future of smart farming with Internet of things." *Journal of Information Technology and Its Applications* 2.1 (2017).

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. Applied Artificial Intelligence, 1–34. https://doi.org/10.1080/08839514.2022.2037254

Jang-Jaccard, Julian, and Surya Nepal. "A Survey of Emerging Threats in Cybersecurity." *Journal of Computer and System Sciences*, vol. 80, no. 5, 2014, pp. 973–993., https://doi.org/10.1016/j.jcss.2014.02.005.

Kanski, Ian. "Center for Advanced Agriculture and Sustainability." *Harrisburg University*, 1 Oct. 2020,https://www.harrisburgu.edu/about/centers-institutes/center-for-advanced-agriculture-and-sustainability/.

Kim, S., Lee, M. and Shin, C. (2018). IoT-Based Strawberry Disease Prediction System for Smart Farming. Sensors, 18(11), p.4051. doi:10.3390/s18114051.

LeClair, Jane, and Sherri W. Ramsay. *Protecting Our Future: Educating a Cybersecurity Workforce*. Hudson Whitman/Excelsior College Press, 2015.

Millett, K., Dos Santos, E., & Millett, P. D. (2019). Cyber-biosecurity risk perceptions in the biotech sector. *Frontiers in bioengineering and biotechnology*, *7*, 136.

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*, 39.

Nchanji, E., Lutomia, C., Chirwa, R., Templer, N., Rubyogo, J., & Onyango, P. (2021). Immediate impacts of COVID-19 pandemic on bean value chain in selected countries in sub-Saharan Africa. Agricultural Systems, 188(0308-521X), 103034. https://doi.org/10.1016/j.agsy.2020.103034

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.

Rejeb, et al."Covid-19 and the food chain? Impacts and future research trends." *LogForum* 16.4 (2020).

Saiz-Rubio, Verónica, and Francisco Rovira-Más. "From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management." *Agronomy*, vol. 10, no. 2, 2020, p. 207., https://doi.org/10.3390/agronomy10020207.

Samori, I. A., Palmer, X. L., Potter, L., & Karahan, S. (2023). Commentary on Biological Assets Cataloging and AI in the Global South. In Proceedings of SAI Intelligent Systems Conference (pp. 734-744). Springer, Cham.

Shahbandeh, M. (22 Aug 2022). U.S. agriculture - statistics & facts. Retrieved from: https://www.statista.com/topics/1126/us-agriculture/#topicHeader__wrapper

Szefer, Jakub, et al. "Cyber Defenses for Physical Attacks and Insider Threats in Cloud Computing." *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, 2014, https://doi.org/10.1145/2590296.2590310.

United Nations (2015). The World Population Prospects: 2015 Revision | Latest Major Publications - United Nations Department of Economic and Social Affairs. [online] Un.org. Available at: https://www.un.org/en/development/desa/publications/world-population-prospects-2015-revision.html.

U.S. Department of Agriculture. (1 Dec 2022). Farming and Farm Income. Retrieved 15 Dec 2022: https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/farming-and-farm-income/

U.S. Department of Agriculture . (2022, June 2). Animal Health Surveillance in the United States. USDA APHIS | Animal Health Surveillance in the United States. Retrieved September 20, 2022, from https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/monitoring-and-surveillance

Van der Burg, Simone, et al. "Ethics of Smart Farming: Current Questions and Directions for Responsible Innovation towards the Future." *NJAS: Wageningen Journal of Life Sciences*, vol. 90-91, no. 1, 2019, pp. 1–10., https://doi.org/10.1016/j.njas.2019.01.001.

Vinatzer, Boris A., et al. "Cyberbiosecurity Challenges of Pathogen Genome Databases." *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019, https://doi.org/10.3389/fbioe.2019.00106.

Whitt, Christine. www.usda.gov. (n.d.). "A Look at America's Family Farms." [online] Available at: https://www.usda.gov/media/blog/2020/01/23/look-americas-family-farms.

Windows, Marc. "Security in Precision Agriculture ." *Diva Portal*, Luleå University of Technology, 2019, http://www.diva-portal.org/smash/get/diva2:1322203/FULLTEXT02.

Wolf, Marilyn. "Cyber-Physical Systems." *High-Performance Embedded Computing*, 2014, pp. 391–413., https://doi.org/10.1016/b978-0-12-410511-9.00008-3.

Yazdinejad, Abbas, et al. "A Review on Security of Smart Farming and Precision Agriculture: Security Aspects, Attacks, Threats and Countermeasures." *Applied Sciences*, vol. 11, no. 16, 2021, p. 7518., https://doi.org/10.3390/app11167518.

Organizational Cybersecurity Post The Pandemic: An Exploration of Remote Working Risks and Mitigation Strategies

Stephen Treacy, Anoop Sabu, Thomas Bond, Joseph O'Sullivan, Jack Sullivan and Peter Sylvester

Department of Business Information Systems, Cork University Business School, Ireland

Stephen.treacy@ucc.ie

Abstract: The Covid-19 pandemic has forced organisations to embrace the largest remote workforce in history, yet this upheaval also brought an increasing number of cyber vulnerabilities to the fore. Organisations must remain committed to not leaving business processes, personal data, or vital infrastructure at risk, which has proved challenging for most. As remote working establishes itself as the new normal, criminals are seeking to capitalize on the widespread cybersecurity uncertainty, and succeeding. Private organisations and cybersecurity professionals must come together to establish robust solutions for home working cybersecurity. *This investigation explores several prevalent cyber risks (private networks, public hotspots, remote desktop protocol, authentication policies, virtual private network configuration and phishing attacks) across three key threat classifications of management, technical and human factors when remote working from the perspective of twenty industry experts. These findings offer key insights to emerging vulnerabilities, while also revealing defined strategies for organisations to help mitigate these challenges.*

Keywords: Cybersecurity, Remote Working, Covid-19, Threat Classification, Phishing

1. Introduction

As the Covid-19 pandemic struck Wuhan, China in 2019 before subsequently spreading across the world in January 2020, businesses had to quickly adapt, abruptly shifting their employees to remote working. As this new way of doing business suddenly became the norm, organisations quickly realized they were being unprecedently challenged to protect valuable data from employee behaviours being targeted by hackers and social engineers. While safe in the comfort of an organizational setting when it comes to cybersecurity, working from home employees tend to develop security amnesia, often abandoning routine security practices, for example establishing authentication procedures, or forwarding suspicious emails, links or attachments to their IT Department. Though employees might initially plan to report these occurrences, including phone calls from social engineers designed to extract valuable information under the guise of pretending to be clients, customers or employees from other offices, they often do not, continuing instead to absentmindedly open links and attachments, and/or engaging in these phone calls without asking for proper verification (Borkovich and Skovira, 2020). Researchers generally agree that unfortunately it is the well-intentioned yet careless worker, vendor, consultant, or other stakeholder that represents as much of a danger to an organisation's cybersecurity as faceless actors on the outside. As a result, valuable lessons that have been learned by organizations in the wake of remote working have often been due to employees abandoning routine security practices when working from home. As more people continue to join and engage with these digital platforms as part of their daily life, so too does the number of cyberattacks that are increasing in many countries, opening new playing fields for cybercriminals to target and exploit. Organisations need to take immediate action to mitigate new cybersecurity risks created by this sudden shift to remote working, because otherwise similar gaps in the organisational and employee protection may be exploited.

2. Literature Review

The Covid-19 pandemic has demonstrated organisations' dependence on information technology, especially the need for adequate cybersecurity to protect the remote workforce and the technologies we are using (Furnell et al., 2021). Cybersecurity can be defined as being the "*collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user's assets*" (ITU, 2008). Modern cybersecurity thus involves detecting behavioural anomalies to prioritise the most severe threats, reducing investigation and threat detection times. According to Shi (2020), it is evident that Covid-19 has impacted cybersecurity spending by firms and organisations, outlining that spending has decreased even though 46% of organisations reported an increased amount of cybersecurity threats related to remote working, with 49% expecting an incident or data breach within a month of the report. Similarly, according to the Federal Bureau of Investigation, the number of successful attacks exploded across the United States by 600%, and across the globe by 300% since the Covid-19 pandemic hit (Borkovich and Skovira, 2020). This represents a massive increase

from the average cost of a data breach in 2019 when it was $3.92 million, or $152 per lost or stolen record and has been directly attributed to the amount of employees now working from remotely due to the invisible threat of Covid-19, a further 1.6% increase from 2018 according to IBM (2019).

2.1 Remote Working Security Risks

Given the significant costs of a data breach to multiple parties, it is critical to implement relevant protection mechanisms in response to general and specific threats. Unfortunately, connecting any system to the internet brings with it inherent security risks as accessing any internet resource opens a communication path that crosses a suite of network equipment from where the connection begins to the destination system, with each communication node representing a potential security vulnerability. For the purpose of this study, the security risks affecting employees engaging in remote working can be divided into three threat classifications: management factors, technical factors, and human factors, as outlined by Wang and Ruan (2020).

Private and public organisations alike are often blamed for cybersecurity breaches due to management's inability to recognize, plan, and fund adequate remote working measures, continuously analyse, scan, test, update and maintain networks and to adequately train employees to recognize and report pretexting cyberattacks, both real and perceived (Bloom, 2014). The difficulty increases therein as employees are now relying on their private networks to gain access to their office network which are not as secure (Ramadan et al., 2021). Improper system configuration has thus become a concern for organisations as wireless networks are more vulnerable than wired networks since intruders can pick up signals from outside the host building allowing the opportunity to breach the network for malicious purposes (Zafft and Agu, 2012). Organisations need to be aware of hackers exploiting weak and unsecured private networks from their employees locations to gain access to their data, as research highlights that many of these networks are not even secured using basic encryption standards (Goldsborough, 2015).

From a technical perspective to counter these issues, mechanisms that reduce network security risks, and use network vulnerability scanning, firewalls, network access controls, and network intrusion detection systems should be implemented (Wang and Ruan, 2020). Unfortunately, an area of concern that has emerged as a network hardware defect has been identified from extant literature as Remote Desktop Protocol (RDP), used to successfully authenticate an unauthorized host that leaves footprints in both the host and network logs. While this service is primarily used by legitimate network administrators, it has also become a primary tool used by attackers since discriminating between legitimate and malicious use of this tool is challenging resulting in the rise of cybercriminals targeting RDP networks (Blanch, 2018). Authentication is similarly vital to maintain the integrity of the system at large, encompassing policies that employees must go through to gain access to confidential and sensitive data. Recent literature has explored how static passwords can increase a user's vulnerability to attack due to their reuse and repeatability making them easier to steal (Channabasava and Kanthimathi, 2019). Most employees choose simple passwords which might lead to a privacy breach or system vulnerability, with Stobert and Biddle (2014) outlining the demands of having multiple accounts on multiple platforms as the reason for password reusability.

Lastly, the human factor represents an additional threat classification as we play an important role in the design, installation, production and maintenance of any infrastructure, with human error leading to the potential disruption or damage of operations (Dhillon and Liu, 2006). This is particularly relevant when the security, stability and reliable operation of the system is critical, with vulnerable virtual private network (VPN) configuration and malicious phishing attacks established as inherent risks. Malicious cyber actors are leveraging existing VPN vulnerabilities to hijack online meetings, teleconferences and online courses which have been established without security controls (Ramadan et al., 2021). Similarly, these same actors have long exploited the human point of failure to gain access to systems, and the number of phishing and cybercrime risks have increased dramatically during the Covid-19 pandemic (Tran, 2020). Phishing attacks target individuals and organisations by collecting personal information and data through different mediums, with these hackers identifying targets and techniques to launch their attacks such as clickjacking, spear phishing, cross-site scripting, man in the middle and malvertising (Alabdan, 2020, Gupta et al., 2018). These schemes are intended to make the person respond without thought, and can be quite difficult to spot, resulting in the human factor being consistently blamed as security's weakest link due to behavioural, social and cultural vulnerabilities (Angwin, 2014).

2.2 Research Gap

Organisations have traditionally spent considerable resources to promote a secure working environment in their offices, however, the Covid-19 pandemic has disrupted this paradigm and forced employees to work remotely. While the workforce can remain just as, if not more productive, organisations need to ensure that these remote working conditions keeps company data as secure as it was in the corporate office (Wallace et al., 2021). Attackers now have multiple entry points and employers must take into consideration the heightened security risks of having a remote and dispersed workforce. Organisations need to be better protected, more secure and should not ignore the vital part their employees play in keeping their data safe (Malecki, 2020). While employees in most cases receive some security training upon joining an organisation, unfortunately this training is rarely delivered on a routine basis, and to date, most organisations have not developed nor circulated a policy and procedure for remote workers (Burke, 2020). We therefore want to take the aforementioned security risks into account from the perspective of industry experts, and state our research questions as: (i) *How are organisations mitigating their management security risks (private network and public hotspot usage) in response to the remote working paradigm?*; (ii) *How are organisations mitigating their technical security risks (RDP and authentication policies) in response to the remote working paradigm*?; and (iii) *How are organisations mitigating their human security risks (VPN configuration and phishing attacks) in response to the remote working paradigm*?

3. Methodology

The purpose of this study is to investigate remote working cybersecurity threats across three key areas from the perspective of security experts, tackling the emerging trends and vulnerabilities organisations are continuing to face while their employees are remote working. To facilitate this investigation, expert judgement can be used to formally bound problems when no data are available through the use of semi-structured interviews (Wilson, 2017). This allowed the researcher to explore emergent topics within the interview setting as they arose, while also empowering the researcher to pursue additional lines of questioning towards cybersecurity threats that the experts had evident experience of. A predominant reason for developing theory from this approach is that it facilitates rich, qualitative evidence, along with testable propositions (Gregor, 2006, Sutton and Straw, 1995). Given how this approach is deeply embedded in rich, empirical data, building theory from evidence gathered produces theory that is accurate, honest, interesting and testable (Eisenhardt and Graebner, 2007). This study uses the definition provided by Garthwaite et al. (2005), who describe experts as being "*persons to whom society and/or peers attribute special knowledge about matters being elicited*" (p. 681). More importantly, it is also the ability to use this knowledge that defines a good expert (O'Hagan et al., 2006). Experts were identified through appropriate case selection methodologies in accordance with Yin (2008) and Seawright and Gerring (2008), ensuring several objectives: (i) a representative sample of expert roles was obtained where similar results were predicted and used as literal replications; (ii) useful variations on the security risks of interest were obtained; and (iii) experts occupied senior roles that made them knowledgeable about the risks being researched. It is from these characteristics, and being consistent with Marshall et al. (2013), that twenty security experts were identified, with titles including directors, founders and managers.

This protocol was aimed to be flexible, and to allow for adjustments based on insights offered through the interviews which was subsequently analysed through the Gioia methodology. This approach facilitates the processing and interpretation of data by establishing concepts and themes, providing an academically rigorous structure with which the data can be interpreted, following five key steps (Gioia et al., 2013): (i) transcript analysis where the researchers searched for similarities and differences among the emerging categories being coded; (ii) first order concepts where the researchers identified if the themes and ideas suggest in the interviews could become conceptual understandings that serve to explain answers to the research questions; (iii) second order concepts where the previous concepts were further broken down to create research findings; (iv) aggregate dimensions where direct quotes are translated through first-order concepts to theoretical and academic understandings of the second order concepts, and are finally reduced to structurally supported and rigorously achieved dimensions that represent findings; and (v) data structures were then created to act as visual aids outlining the steps involved from inception of the theme to the themes created in the aggregate dimensions.

4. Findings

4.1 Research Question 1

The issue of insecure public and private Wi-Fi networks was highlighted as being a problem area for the experts in this investigation. Employees were advised to use secured and private Wi-Fi when working remotely, and on occasion when an employee needs to connect over a public hotspot, they were advised to utilise a virtual private network (VPN). The practice of using public hotspots is one that should be discouraged: "*Organisations can't use public or Wi-Fi, or hotspots*" (E-11). IT security governance teams were encouraged to "*explicitly outline an organisation's stance on the use of public and private Wi-Fi to prevent ambiguity*" (E-3). Organisations should grant network access to an employee's home IP address, and no access should be given to an IP address deemed potentially unsafe. Even with tools and policies organisations may have in place, the experts advised to whitelist and blacklist IP addresses to minimize the security risk level. In addition, since the use of public hotspots by employees "*can further expose an organisation's network and make it even more susceptible to attack or infiltration*" (E-8), strict policies such as IP filtering have become even more vital for organisations. The majority of experts agreed that effective training is necessary to promote awareness of network vulnerabilities when working from home: "*Employees need to be regularly trained on the differences between public and private Wi-Fi, especially on the associated risks*" (E-14), at least annually. Some experts, however, were conflicted, arguing that "*training does not generally work*" (E-7), and that for some organisations there is a need to focus on employee behaviour rather than policy. This lack of confidence in employees results in a zero-trust approach where "*the underlying network should not be trusted… (as it) exposes organisations to other security threats*" (E-8). Private Wi-Fi should also be encrypted, with authentication protocols being enforced for organisations to have "*Wi-Fi enabled entrusted devices*" (E-14), with all devices and technologies being updated and patched regularly: "*IT teams are able to push security patches and updates remotely*" (E-19), often with "*proprietary software*" (E-1). The experts also stressed to maintain zero-trust policies "*irrespective of remote work or non-remote work*" (E-3).

Experts highlighted the need for comprehensive governance strategies in the event of a breach through these networks, including the use of response teams, data security, and backups, along with the need for "*a full governance model from notification of a breach to first responder… (and) an escalation strategy from our IT team, governance and security team, through to our response team*" (E-3). Organisations were also advised to have both internal and external communication plans to issue in the event of an attack. These can be communicated appropriately to both team members and customers should the system need to go offline for any period while a solution is applied. The experts described how some large organisations also have dedicated security operation centers who review all possible breaches and "*if there is a breach of any kind, we immediately reach out to the relevant stakeholders*" (E-13). The team then revokes all access to the network that other employees might have. Furthermore, organisations should have data security policies that cover the event of breaches, where employees will immediately report "*the instance of a breach*" (E-4) and be well educated on what steps should be pursued thereafter. Organisations were advised to have cloud backups or replicas of the main systems to prevent data loss, with the ideal response strategy being able "*to have an exact replica of their systems that is totally isolated and decoupled from the production systems*" (E-8), as the attackers may still be operating within the network, infiltrating more data and infecting other machines resulting in everything needed to be shut down. This strategy aids the organization in shutting down existing machines which are running, allowing the response teams to perform analytics and investigate the nature of the attack: "*if its ransomware, it will spread and make the data breach worse*" (E-8).

4.2 Research Question 2

Similarly, the pandemic forced organisations to rely on RDP for business continuity, with the experts arguing that RDP can only be utilized by organisations if it is managed properly with certain controls. For instance, one expert advised organisations to "*limit RDP software solely to within the local area network (LAN)*" (E-14). Other experts described how they prevent employees downloading RDP software, and how organisations have RDP sites blocked from being accessible: "*we're able to block standard RDP ports from being used on the network*" (E-6). Any other use of RDP should not be permitted. Unfortunately, the consensus among experts was that RDP is suspect to unsecure and outdated technology: "*RDP is very, very weak for security. We don't do it, we can't do it, its forbidden*." (E-9), with one expert even going so far as to describe RDP as being "*a honeypot for attackers*" (E-8). The security risks were highlighted as being too great as the technology was described as easily penetrable and largely outdated. Experts stressed that organisations should not invest on securing RDP, but to

spend money replacing it with more modern solutions: "*The only discussion around RDP should be how to we replace the existing service so that we don't use RDP anymore*?" (E-8). For organisations that continue to use RDP however, several recommendations were presented by the experts, including the adoption of monitorization tools, VPNs, and encryption protocols. Additionally, organisations have also used VPN connections as a solution to mitigate RDPs vulnerabilities, although "*not for any work deemed valuable or confidential*" (E-12). While RDP has been widely documented by the experts as being a vulnerable technology and advocating against its use, organisations continue to use it internally necessitating response strategies due to high risk of attack. The experts described how secure cloud computing, employee use of RDP and quarantining of servers could play an important role in responding to threats of this nature. Experts recommended organisations to ensure adequate measures and response strategies are in place for after any attack, as "*RDP is a huge headache in terms of online security, and the vulnerabilities and issues that come with it*" (E-6). Many organisations the experts worked with either had RDP banned, or only allowed their IT teams to use the technology. Experts described this justification as the more technically qualified the employee, the quicker a data breach or infiltration over RDP can be discovered. The average employee will not be able to identify a breach, meaning infiltration goes unnoticed for longer, causing increasing levels of fallout: "*If you are getting an attack on RDP, that means the hackers are already in your network. To get to the RDP, you have to first breach the cloud network, but if that's secure enough, its prevented*" (E-1). An effective response strategy should also detail when a server should be "*quarantined*" (E-16) to provide extra protection when a server is infiltrated.

Authentication was also discussed as an important process to allow employees gain access to confidential and sensitive data, while dissuading malicious actors from breaching their networks. As organisations initially struggled with remote working protocols, the topic of IT security and governance policies were outlined by experts as having a huge influence on how authentication password policies were implemented across all devices. There are several layers of authentication that organisations need to follow, ranging from an employee's own workstation, to VPN and single sign on, with experts arguing that improvement is needed across all areas, including: "*devices should not be jailbroken, or rooted, devices should have a passcode, copy/paste is disabled, screenshot is disabled, devices should contact mobile device management every seven days, no VPN can be configured on personal devices...*" (E-17). Experts also advised organisations to adopt multi-level authentication, that should initially seek to "*authenticate the workstation via a certificate exchange between computer and VPN server. Next phase is your traditional username and password, followed by multifactor hardware or software tokens*" (E-12) which should consist of a "*password of alphanumeric values and authentication code from authenticator app*" (E-20). Experts recommended that all passwords should be changed every three months to add an extra layer of rigor. These steps can help provide a rigorous authentication policy to mitigate the possibility of cyber-attack or infiltration. Unfortunately, experts revealed that many organisations are using the same level of authentication policies they were using pre-pandemic. In those cases where infiltration does occur, a proactive response is required across several key areas with experts providing several examples of strategies: "*Firstly, disable the user account immediately. Secondly, close all user sessions. Thirdly, create a script so that service desk engineers can run it with a click of a button to perform all activities including password changes, then notify the organisation's security team*." (E-17). The experts argued that communicating updates to authentication software should be the last critical step to take: "*You don't have to go to Microsoft to download the latest updates for example, it is being pushed by corporate*" (E-9). This results in all technology being up to date with the latest security updates and patches, with organisation's information security officers communicating these protocols and response strategies with employees. Red team tests were also described as another way where organisations can test portals of systems, identifying what threats devices are exposed to: "*You can look at what devices are exposed to authenticate which have a certificate with the company name… You have to keep doing that scan at least two or three times a day on your public facing infrastructure to contain any inadvertent exposures*" (E-5). Experts outlined the necessity of being extremely aggressive in finding exposures before attackers do, because attackers are always looking.

4.3 Research Question 3

The experts agreed VPNs helped organisations during the initial seismic shift to remote working, scaling up their infrastructures to handle the large amounts of traffic passing through their system from employees working virtually, specifically through increasing VPN encryption and authentication. VPN connectivity must also be managed on a layered basis, which provides added security across three factors of authentication: "*We authenticate the machine, the user, and then we use a form of multi-factor as well on top of that. It would be incredibly difficult for someone to connect who does not have authorization*" (E-12). Every expert stressed how critical employee cybersecurity training is to their organization, with the use of VPN being a prominent factor in

that training: "*During the Covid-19 pandemic, we strengthened our VPN and were more vocal about when you should use it, and when you should not*" (E-15). Some organisations tested their employees four times a year on the evolving risks as part of a "*quarterly certification*" (E-3) ensuring that their employees knew how to use VPNs effectively. Company policies have been forced to make wholesale changes and are now being reviewed and updated annually. For example, organisations now have factored in "*audit requirements*" (E-12) as part of their annual training and policy reviews, with other organisations introducing a "*split tunnelling*" (E-17) VPN system that enables only office traffic to pass through a VPN and the public connection going through a user's internet connection. These controls help mitigate evolving cybersecurity issues and demonstrate how important it is for companies to regularly review their policies. While devices that use VPN connectivity can be breached, they are more secure than a network accessed without a VPN, with experts outlining several mechanisms with which to monitor VPN use, for example, monitoring any "*land-speed violations*" (E-5). This mechanism involves the monitoring of logins, identifying the geographical location on where access points are coming from, and if the user is coming from various countries in a manner of minutes. All "*anomalous activity*" is immediately investigated by the security team to ensure there are no compromises on the network. Layering the infrastructure allows for more concise response strategies after identifying a breach, depending on what environment may have been accessed by external actors, for example whether access was gained to "*the development environment, or the operations environment etc…*" (E-10). Once the breach is isolated, forensic diagnostics can be subsequently implemented to measure the impact. These response strategies center around continuous monitoring and investigating to counter the risks associated with data breaches.

Experts also described phishing attacks as being the most prevalent form of cyberattack since the pandemic began and after organisations transitioned to remote working. Given their constantly evolving nature, experts highlighted the importance of training, monitoring, management, and decoy methods to combat this issue. Some organisations aimed to provide this training on a quarterly basis, but at the outset of the pandemic they "*were doing it nearly weekly, as phishing is the most prevalent type of attack to the employee base*" (E-14), with other experts claiming that "*You can never do enough education in this space*" (E-15). Several experts described how these organisations manage the risk of phishing by performing regular tests on their employees: "*We do phishing campaigns at least once a quarter*" (E-15), whereby they send out fake phishing emails to their employees, and receive a report on who clicked into the mail: "*Our team sends everyone a link, and tell them* "You all need new computers, please go on this link and choose which one you want." *A lot of people click through because they think it's from someone they know, which it is, but the link is often dangerous*" (E-7). The employees that accessed the content would then receive links to more education on phishing. Organisations are also increasing their reliance on anti-phishing software to survey all inbound emails, with some software packages with built-in functionality to identify phishing attempts, capable of catching "*millions of phishing items every day*" (E-15), however, the more sophisticated the attack, the more likely it might slip through the filtering technology. Larger organisations were described as having the ability to dedicate entire teams during the Covid-19 pandemic to make monitoring and management of phishing their full-time job due to the surge of attacks. Communication was identified as a vital component to any occurrence of phishing, identifying possible and real issues and making their stakeholders aware of any phishing campaigns. Reporting any suspicious activity by employees also aids organisations: "*We have a phishing button on our email clients that enables people to report back into security and say something isn't right, can we investigate*" (E-14). This type of reporting means that phishing emails that successfully slip through email-filters can be quickly dealt with. If a phishing email is clicked on by an employee, the experts described how organisations have protocols in place to track back and identify the impact path. While some operating procedures can vary, it is critical at minimum to contact the employee immediately, assess what was clicked on and if anything was shared, reset passwords, and shut down all active sessions of the user.

5. Discussion and Conclusions

Our study distinguishes itself from prior cybersecurity research in terms of literature selection criteria, research focus, and research output. This study identifies several approaches on how to mitigate cyber-attacks in these areas to maintain an optimal remote working environment. Firstly, from a management perspective, these findings reveal several regulatory and training recommendations that should be implemented within the technical architecture of an organization, for example, effective use of data policies and encryption methods should be developed to ensure cyber safety of employees working on their private networks. Additionally, security protocols should be enforced on users such as setting up router passwords, hiding Service Set Identifiers (SSID), keeping the system firewall updated, using efficient antivirus software and other protective tools like IP address filtering along with whitelisting of IPs to prevent involuntary/voluntary access to malicious websites.

The findings also highlight the emphasis that should be placed on utilising VPN, whether on private or public networks to have an additional layer of security. For public networks specifically, training must be provided on safe usage and to develop a zero-trust approach. Secondly, from a technical perspective, RDP was unanimously identified as a technology that needs to be replaced due to the inherent security vulnerabilities, and to ensure multi-factor authentication is adopted as a basic requirement. While RDP is still in use in many organisations, companies are now migrating to more reliable technologies like VPN and Citrix to deliver secure remote working environments. The findings also highlight how MFA security strengthens the encryption architecture irrespective of what the primary layer of encryption is, which could differ across companies. The primary form of encryption however should not be stagnant, as expressed by experts and proven methods that passwords/access policies need constant updating to stay ahead of potentially malicious actors. Thirdly, from the human perspective, this can be viewed as one of the most prominent causes of breaches irrespective of the security measures implemented. Blindly trusting that employees will do the right thing can lead to catastrophic consequences. To that end, technology and employee training needs to be coupled together to maximize security efficiency. With the surge of VPN usage due to the sudden shift to remote working, organisations have had to significantly increase their infrastructure capacities to ensure a smooth and secure transition, however, this should not come at the expense of security, as there has to be extensive work put into the VPN client in the first place before it is distributed to the employees. Similarly, phishing emerged as a significant human threat level event, that could bypass all security measures in place and cause havoc to an organisation. The findings recommend developing a combination of phishing filtering technology and employee awareness training to avoid successful phishing attacks, in addition to frequent tests and monitoring systems. The findings presented herein should be considered as a roadmap with which the organisation can use to develop efficient mitigation strategies for remote working threats. Indeed, even providing visibility over these areas allows organisations to critically analyse their needs and capabilities to further refine existing security policies. These findings reveal key insights to develop new mitigation strategies, technologies and flexible solutions for remote working, and methods to facilitate optimal stakeholder engagement in their delivery that aligns with business needs. While we endeavoured to achieve the highest levels of rigour and accuracy, as is true of any research this study has several limitations, which can be addressed by future research. One limitation is the fact that only six technologies were included in the risk categorisation, based on relevance and priority, from the wide array of technologies that are an aspect of the remote working environment. Topics that have not been studied may become more of a priority later on as the current environment evolves; leaving a gap to build on our current findings. In addition to advancing our understanding of cybersecurity practices within organisations, this investigation can also inform policy makers of the possible determinants and outcomes of these specific cybersecurity incidents. In doing so, it may also have important implications for improving the discussion of cybersecurity activities in financial disclosures, and effective communication of such information with relevant stakeholders. We are still at an early stage of this transformation however, which presents future investigations with an exciting research area.

References

ALABDAN, R. 2020. Phishing attacks survey: types, vectors, and technical approaches. *Future Internet,* 12**,** 168.

ANGWIN, J. 2014. Dragnet Nation: A Quest for Privacy, Security, and Freedom in a World of Relentless Surveillance: Chapter 1: Hacked. *Colo. Tech. LJ,* 12**,** 291.

BLANCH. 2018. *Five issues facing secure remote access to IIoT machines* [Online]. Machine Design. Available: https://www.machinedesign.com/automation-iiot/article/21836802/five-issues-facing-secure-remote-access-to-iiot-machines [Accessed 21/10/2021].

BLOOM, N. 2014. To raise productivity, let more employees work from home. *Harvard Business Review, January–February*.

BORKOVICH, D. J. & SKOVIRA, R. J. 2020. Working from home: Cybersecurity in the age of COVID-19. *Issues in Information Systems,* 21..

BURKE, S. 2020. *Coronavirus Is Creating A Global 'Work-At-Home' Culture* [Online]. CRN. Available: https://www.crn.com/news/cloud/coronavirus-is-creating-a-global-work-at-home-culture [Accessed 15/11/21].

CHANNABASAVA, H. & KANTHIMATHI, S. Dynamic Password Protocol for User Authentication. Intelligent Computing-Proceedings of the Computing Conference, 2019. Springer, 597-611.

DHILLON, B. & LIU, Y. 2006. Human error in maintenance: a review. *Journal of quality in maintenance engineering*.

EISENHARDT, K. M. & GRAEBNER, M. E. 2007. Theory building from cases: opportunities and challenges. *Academy of Management Review,* 50**,** 25-32.

FURNELL, S., HANEY, J. & THEOFANOS, M. 2021. Pandemic Parallels: What Can Cybersecurity Learn From COVID-19? *Computer,* 54**,** 68-72.

GARTHWAITE, P. H., KADANE, J. B. & O'HAGAN, A. 2005. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association,* 100**,** 680-701.

GIOIA, D. A., CORLEY, K. G. & HAMILTON, A. L. 2013. Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational research methods,* 16**,** 15-31.

GOLDSBOROUGH, R. 2015. Staying Safe When Using Wi-Fi. *Teacher Librarian,* 42**,** 65.

GREGOR, S. 2006. The Nature of Theory in Information Systems. *MIS Quarterly,* 30**,** 611 - 642

GUPTA, B. B., ARACHCHILAGE, N. A. & PSANNIS, K. E. 2018. Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems,* 67**,** 247-267.

IBM. 2019. *Cost of a data breach study: Global overview* [Online]. Available: https://www.ibm.com/downloads/cas/ZBZLY7KL [Accessed].

ITU. 2008. *Definition of cybersecurity referring to ITU-T X.1205, overview of cybersecurity* [Online]. Available: https://www.itu.int/en/ITU-T/studygroups/com17/Pages/cybersecurity.aspx [Accessed 13/10/2021].

MALECKI, F. 2020. Overcoming the security risks of remote working. *Computer Fraud & Security,* 2020**,** 10-12.

MARSHALL, B., CARDON, P., PODDAR, A. & FONTENOT, R. 2013. Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of computer information systems,* 54**,** 11-22.

O'HAGAN, A., BUCK, C. E., DANESHKHAH, A., EISER, J. R., GARTHWAITE, P. H., JENKINSON, D. J., OAKLEY, J. E. & RAKOW, T. 2006. *Uncertain judgements: eliciting experts' probabilities*, John Wiley & Sons.

RAMADAN, R. A., ABOSHOSHA, B. W., ALSHUDUKHI, J. S., ALZAHRANI, A. J., EL-SAYED, A. & DESSOUKY, M. M. 2021. Cybersecurity and Countermeasures at the Time of Pandemic. *Journal of Advanced Transportation,* 2021.

SEAWRIGHT, J. & GERRING, J. 2008. Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly,* 61**,** 294-308.

SECURITY. 2020. *Half of organisations experienced security incidents while working remotely* [Online]. Security Magazine. Available: https://www.securitymagazine.com/articles/93505-half-of-organizations-experienced-security-incidents-while-working-remotely [Accessed 15/11/2021].

SHI, F. 2020. Surge in security concerns due to remote working during COVID-19 crisis. Barracuda.

STOBERT, E. & BIDDLE, R. The password life cycle: user behaviour in managing passwords. 10th Symposium On Usable Privacy and Security ({SOUPS} 2014), 2014. 243-255.

SUTTON, R. I. & STRAW, B. M. 1995. What theory is not. *Administrative Science Quarterly,* 40**,** 371-384.

TRAN, C. 2020. Recommendations for ordinary users from mitigating phishing and cybercrime risks during COVID-19 pandemic. *arXiv: 2006.11929 v1*.

WALLACE, S., GREEN, K., JOHNSON, C., COOPER, J. & GILSTRAP, C. 2021. An Extended TOE Framework for Cybersecurity Adoption Decisions. *Communications of the Association for Information Systems,* 47**,** 51.

WANG, Z. & RUAN, Q. 2020. Research on network security subsystem based on digital signal. *Journal of Intelligent & Fuzzy Systems,* 38**,** 97-103.

WILSON, K. J. 2017. An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting,* 33**,** 325-336.

YIN, R. 2008. *Case Study Research: Design and Methods 4th Edition*, London: Sage Publisher.

ZAFFT, A. & AGU, E. Malicious Wi-Fi networks: A first look. 37th Annual IEEE Conference on Local Computer Networks-Workshops, 2012. IEEE, 1038-1043.

A Quantitative Risk Assessment Framework for the Cybersecurity of Networked Medical Devices

Maureen S. Van Devender and Jeffrey Todd McDonald
University of South Alabama, Mobile, AL USA

mvandevender@southalabama.edu
jtmcdonald@southalabama.edu

Abstract: Medical devices are increasingly the source of cybersecurity exposure in healthcare organizations. Research and media reports demonstrate that the exploitation of cybersecurity vulnerabilities can have significant adverse impacts ranging from the exposure of sensitive and personally identifiable patient information to compromising the integrity and availability of clinical care. The results can include identity theft and negative health consequences, including loss of life. Assessing the risk posed by medical devices can provide healthcare organizations with information to prioritize mitigation efforts. However, producing accurate risk assessments in environments with both sparse historical data and a lack of validation regarding the accuracy of forecasts is particularly challenging. We present a risk assessment framework for quantifying the risk posed by connected medical devices in trusted healthcare networks. Our framework is built upon prominent existing frameworks and guidance for general risk assessment and cybersecurity risk assessment. We add a method for quantifying risk, which to our knowledge is novel in the context of medical devices on trusted networks. The framework provides a structure for combining publicly available information along with expert elicitation about threats, vulnerabilities, and consequences. The goal is to provide healthcare organizations with actionable information for prioritizing and mitigating risks in medical devices.

Keywords: Risk assessment framework, medical device security risk, threat/vulnerability/asset (TVA) models

1. Introduction

Cybersecurity incidents are on the rise, and organizations face the challenge of protecting against attacks from an adversary that is increasing in sophistication (Verizon, 2022). Healthcare organizations are experiencing significant security incidents (HIMSS, 2022) (Williams and Woodward, 2015). Research has demonstrated that medical devices pose cybersecurity vulnerabilities in healthcare networks and that they are being used as key pivot points by attackers to establish command and control within networks from where data can be exfiltrated and ransomware may be launched (TrapX Labs, 2015) (Federal Bureau of Investigation (FBI), 2022). Email phishing attacks are reported as a significant source of cybersecurity exposure (HIMSS, 2022), and while antivirus protection may quickly clear malware from workstations, the malware may swiftly spread to medical devices where they are not as well protected (Reel and Robertson, 2015). Once on an unprotected medical device, malicious actors can investigate network resources and plan their attack. In addition, while Internet of Medical Things (IoMT) devices add value to healthcare delivery, they also present cybersecurity challenges to healthcare organizations (Hireche *et al.*, 2022).

The effect of cyberattacks in healthcare can be the disruption of information technology operations, the unavailability of clinical care, or damage to systems and devices (HIMSS, 2022). The extent of adverse impacts to patient health due to cybersecurity events is largely unknown due to a lack of mechanisms to examine patient safety in the context of cybersecurity (US Dept of Homeland Security, 2019). However, two legal proceedings alleging deaths related to ransomware attacks on hospital networks have been reported (Collier, n.d.; Eddy and Perlroth, 2020).

Reports that medical devices are at increased risk of cybersecurity vulnerabilities, evidence signifying increased attacks in healthcare environments, and claimed vulnerabilities in specific medical devices prompted interest in defining a framework for conducting risk assessment specific to medical devices. We present a risk assessment framework for networked medical devices that can serve as input into overall risk management.

Section 2 provides related work to our approach and section 3 proposes a risk assessment framework for networked medical devices. In section 4 we conclude with the merits and limitations of the proposed framework and discuss future work.

2. Related Work

Lee et al., (Lee *et al.*, 2012) examine the challenges and research directions in Medical Cyber Physical Systems (MCPS). The authors identify the increase in the interoperability of medical devices as providing advantages and improvements in healthcare delivery, while also creating greater attack surfaces. They state that it is essential

interoperable medical devices be secure for the primary reasons of their propensity to be deployed in life critical situations and to have access to sensitive health information. They conclude that the domain of MCPS provides a unique set of challenges that are distinct from other cyber physical systems. While they identify cybersecurity challenges in networked medical devices, they do not propose a solution for assessing the risks.

Sappal and Prowse [28] propose a method for lifecycle management of connected medical devices that emulates electromechanical preventative maintenance and technology management corrective maintenance practices that are already established in healthcare organizations. It provides for lifecycle management of connected medical devices through tracking, scoring, and reporting on cybersecurity vulnerabilities by medical devices. While their approach does not attempt to assess or quantify risk, it provides a means to prioritize vulnerabilities by a weighted average that includes device function, location, operating system, a medical device CVSS score [29], and the failure consequence.

Kaplan and Garrick (Kaplan and Garrick, 1981) describe risk as in terms of an overall risk triplet – threat, vulnerability, and consequence. The triplet addresses respectively what can happen, how likely it is to happen, and what are the consequences if it does happen. Kaplan and Garrick demonstrate that probability of frequency aligns with the Bayesian approach. Our framework adopts this conceptual view of risk and relies upon it in our definition of risk.

The domain of cybersecurity risk is one that lacks historical data on which to predict future outcomes (Hubbard and Seiersen, 2016). Eliciting the judgment of experts has been used to support risk estimation in domains where there is little historical data on which to predict outcomes (Cooke and Goossens, n.d.; Ortiz *et al.*, 1991). Krisper, et al. (Krisper *et al.*, 2020) demonstrate a process of using multiple experts and combining their judgments using a weighted average based on their performance in earlier calibration tests with one cybersecurity risk scenario.

Lichtenstein and Fischoff (Lichtenstein and Fischhoff, 1980) demonstrate that training experts to improve probability assessments can be an effective means for improving their accuracy. They demonstrate a process of eliciting subjective probabilities from the experts and providing immediate feedback on their performance. Practicing this training technique provides an environment that supports improvement in probability estimates with minimal training. These techniques have been used across domains including nuclear energy (Goossens *et al.*, 2008) and conservation science (Martin *et al.*, 2012). This research demonstrates the value of calibration in improving expert judgment.

Pardue et al. [37] developed the foundational database-driven approach to risk assessment upon which this research is built. The research method was a proof of concept using a hypothetical scenario in the healthcare domain. Pardue et al. underpin their work by identifying the essential elements for information security assessment as Threat, Vulnerability, Asset (TVA) from the work of Hoffman, et al. [38] and Whitman [39] as the core structure for their design. We build upon the structure provided in this research and add quantification of risk.

Previous risk assessment research has identified security challenges in medical devices, proposed solutions for vulnerability management, applied expert judgment to risk assessment, and proposed solutions. However, there is minimal empirical research investing a cybersecurity risk assessment framework that provides a quantified estimate of the expected loss related to the exploitation of vulnerabilities specific to medical devices. We propose a framework that provides for the identification of risk scenarios considering published cybersecurity vulnerabilities, the identification of threat actors, and the estimation of impact using expert elicitation and weighted criteria that serves to reduce the uncertainty associated with the impact of risk scenarios. The risk assessment provides quantified estimates of risk scenario magnitudes that can be used to prioritize risk mitigation efforts and serve as input to an overall risk management program.

3. Framework for Risk Assessment of Networked Medical Devices

In this research, risk is defined as a measure of the extent to which the organization is threatened by a circumstance or event, expressed as a function of the adverse impact of the circumstance or event and the frequency of its occurrence (1) :

$$R_e = F_e I_e \quad (1)$$

In our framework we refer to circumstances or events as risk scenarios. It is not uncommon for risk to be a function of likelihood and impact. We use frequency instead of likelihood for its suitability for quantifying risk over a given time period.

Existing risk assessment frameworks were investigated to gain insight into methodologies. While any of a number of frameworks could have been chosen as guides to developing a medical device risk assessment framework (HITRUST Alliance, 2019) (CIS, 2019; ISO, n.d.) (Caralli *et al.*, 2007), several sources were selected for their suitability to this context. First, the Framework for Improving Critical Infrastructure Cybersecurity (Cybersecurity Framework) (NIST - CF, 2018) was selected as the overall guide for conducting the risk assessment. The choice for the Cybersecurity Framework was based on the prevalence of this framework's use in the healthcare sector (HIMSS, 2018) and its use in conducting risk assessments at our partnering healthcare facility. In addition, the approach to risk assessment in the Cybersecurity Framework is intended to be consistent with the approaches described in the ISO/IEC standards (ISO, n.d.), a prominent risk assessment tool set. Second, the NIST SP 800-30 Guide for Conducting Risk Assessment (Ross, 2012) is used for the stepwise structure it provides. Third, the Factor Analysis of Information Risk (FAIR) (Freund and Jones, 2014) framework is relied upon in this research in the process of identifying assets, assessing threat actor capabilities, threat event frequency, loss event magnitude, and in quantifying risk. FAIR provides structure and detail in areas where the Cybersecurity Framework and NIST SP 800-30 are more general.

The risk assessment function within the Cybersecurity Framework includes identifying vulnerabilities, receiving threat intelligence from information sharing forums, identifying internal and external threats, identifying potential impacts and likelihoods, and combining threats, vulnerabilities, and likelihoods to determine risk. Our framework, based on the four-step process of the NIST SP 800-30 Risk Assessment Process is shown in **Figure 1**. A description of the execution of each of the steps in this research follows.

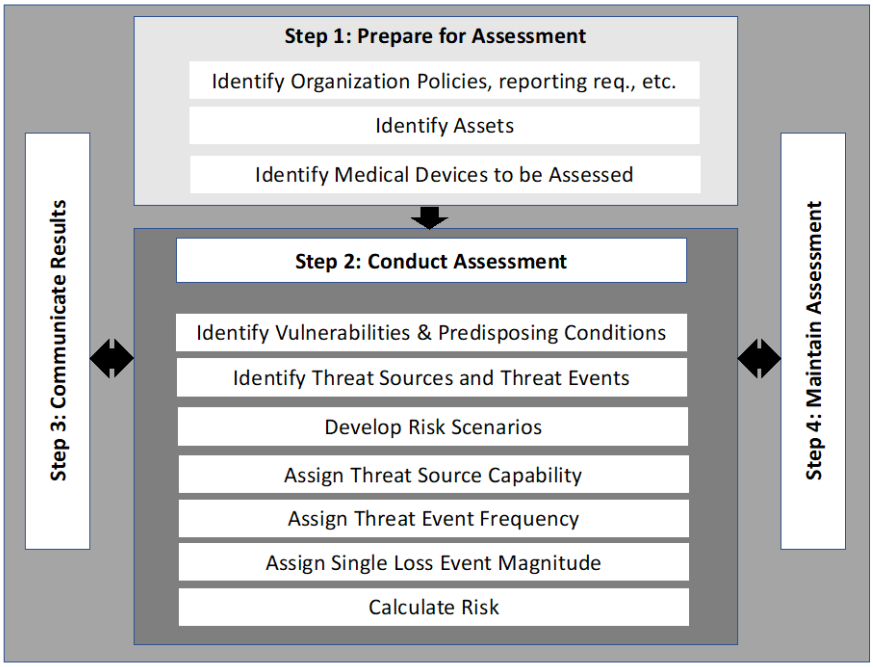


Figure 1: Medical Device Risk Assessment Framework

3.1 Step 1: Prepare for Assessment

The purpose of preparing for the risk assessment is to establish the context and scope of the assessment. This may include identifying organizational policies and requirements for the risk assessment and identifying reporting requirements and methodologies to be used. While this will vary among organizations, identifying the assets, the medical devices, and the timeframe for the assessment is essential to preparing for the risk assessment of medical devices.

3.1.1 Identify Assets

In a medical device risk assessment, it would be logical to consider the medical devices to be the asset. However, Landoll (2011) defines assets in risk assessment as those items considered valuable to an organization and its stakeholders. Examples of assets in the context of a healthcare organization may be patient safety, Protected Health Information (PHI), and business revenue. In reflecting on the organization's assets to be protected based

on FAIR guidance (Freund and Jones, 2014), it became clear that the medical devices themselves are not the assets of value to the organization and its stakeholders. Rather, medical devices may contain assets or be paths to assets.

The process of identifying the assets to be protected involves identifying the things of value to the organization and its stakeholders that could risk well-being if they were to be lost or damaged. Asset identification and asset valuation is obtained by eliciting experts such as organization leadership like finance leaders and executives. In addition to the identification of the assets, an asset valuation can be provided by these experts. Valuation of the assets helps to estimate the potential impact of a risk scenario and to evaluate appropriate controls (Landoll, 2011).

3.1.2 Identify Target Devices

The OCTAVE Allegro framework (Caralli *et al.*, 2007) describes information asset containers as places where information assets are stored. Medical devices may be containers for information assets, and they may also be paths to assets. We conclude that medical devices can be in a category that is analogous to the OCTAVE information asset container. We therefore develop a component of the framework for medical devices that is separate from the assets to be protected. We refer to this component as Target Devices.

The target devices selected for a risk assessment are identified by the organization in the preparation step as identified in NIST SP 800-30. The preparation step precedes the risk assessment therefore the process of selecting the medical devices is not discussed in detail here. The device selection could be all or a subset of networked medical devices.

Once device selection is made, an inventory of medical devices is examined and curated to include all the following that apply: asset model number, operating system and version, firmware version, and all installed software with its version. Collecting all of this information may present a challenge for some devices because visibility into the components of medical devices is not always possible and is an ongoing challenge (NTIA, 2019). With this limitation, every effort should be made to identify the components of each device.

The Security Content Automation Protocol (SCAP) (NIST - SCAP, n.d.) is a community effort overseen by NIST that contains standardized expressions. Of interest to us in SCAP is Common Platform Enumeration (CPE) (NIST - CPE, n.d.) which is a standard method for identifying hardware, software, and operating systems. NIST hosts and maintains the official CPE dictionary, which is available to the public. A search of the CPE dictionary reveals that it contains entries for medical devices, including component specifications that have known vulnerabilities in the NVD. Because CPE follows a rules-based nomenclature, the CPE can be derived for each medical device and its components. This can then be used to search the NVD for vulnerabilities. We identify the CPE for each medical device and its components to assist later in the identification of published vulnerabilities.

3.2 Step 2: Conduct Risk Assessment

The risk assessment can begin once preparation is complete. Our framework is composed of the identification the risk triplet of threat, vulnerability, and consequence (Kaplan and Garrick, 1981) that can negatively impact the organization and its stakeholders through damaging the value of assets. The risk assessment tasks are: identify vulnerabilities and predisposing conditions, identify threat sources and events, develop risk scenarios, assign threat source capability, assign threat event frequency, assign single-loss event magnitude, and calculate risk as the expected loss magnitude. Each task is described here.

3.2.1 Identify Vulnerabilities and Predisposing Conditions

Vulnerabilities associated with each of the devices that are reported by the manufacturer should be logged and used in the risk assessment process. In addition, Open Source Intelligence (OSINT) (Hassan and Hijazi, 2018) techniques can be used to identify cybersecurity vulnerabilities that may have been reported by other parties.

Common Vulnerabilities and Exposures (CVEs) (NIST- NVD, n.d.) is a program under the direction of the MITRE Corporation for the purpose of maintaining a list of publicly known cybersecurity vulnerabilities. CVE is intended to be a comprehensive catalog of publicly disclosed cybersecurity vulnerabilities. The NVD is a publicly searchable database that contains each CVE along with some additional information. It includes a textual description of the vulnerability that may be useful in characterizing the risk scenario(s) that could result in an exploitation. NVD also provides links to external information about the vulnerability. Attributes useful to risk assessment are provided in the NVD, such as a list of all affected hardware and software using CPEs (NIST - CPE,

n.d.). The CPE is helpful in identifying exactly which computing components are affected by the CVE. In step 1, preparation, we established CPEs associated with each target device to facilitate searching the NVD. A Common Vulnerability Scoring System (CVSS) base score (FIRST, 2019) is assigned to each vulnerability in the NVD.

The CVSS base score is composed of two sets of metrics – exploitability metrics and impact metrics, and it provides qualitative and quantitative values. While the CVSS score is not intended to be a measure of risk, there are several metrics in the base score that can be useful in characterizing risk. The metrics we identify to be useful are the Exploitability sub score and the set of impact metrics - Confidentiality, Integrity, and Availability (CIA) Impacts. The exploitability sub score indicates the ease and technical means by which a vulnerability can be exploited. The CIA impact metrics reflect the consequences of a successful exploitation.

We use the CIA triad as security effect attributes in our estimation of adverse impacts. We define security effect attributes as those that effect the systems and/or information assets of the organization. The CIA triad is a widely accept model for information security (Samonas and Coss, 2014). Our impact assessment considers the potential compromise of one or more of the CIA components.

Predisposing conditions are those conditions that could contribute to the likelihood that one or more threat events would result in negative consequences (NIST, n.d.). Conditions of interest for each device include: whether it contains PHI, the physical security status of the device, the FDA class (FDA, 2018) of the device, if it is receiving software and firmware updates, and if the device has been designated as 'end of life' or no longer supported with updates by the manufacturer. **Table 1** shows example predisposing conditions of interest in risk assessment. These attributes are examples of those that can be useful in understanding the riskiness of the device.

Table 1: Medical Device Predisposing Condition Attributes

Medical Device	Physical Status	FDA Class	PHI?	Getting Updated?	End of Life?
Device 1	In secure room; user authentication required	3	Y	Y	N
Device 2	user authentication required	2	N	Y	N
…	..				
Device n	mobile; authentication required	2	Y	N	Y

3.2.2 Identify Threat Sources and Threat Events

This research considers threats that are specific to the medical devices, described as the information system level in NIST SP 800-30 (Ross, 2012). In this assessment, threats are decomposed into threat sources and threat events.

3.2.3 Identify Threat Sources

Threat sources are characterized as the intent or method targeted at a vulnerability, or a situation and method that may accidentally exploit a vulnerability (Ross, 2012). The definition of a threat source used in this research is anything that is capable of acting in a manner that can result in harm (Freund and Jones, 2014). Threats sources are generalized as: adversarial/malicious and human errors of omission or commission (Ross, 2012). They are identified as groups rather than individuals. For example, a threat actor may be a malicious insider, but not a specific individual within the organization.

Threat sources per NIST SP 800-30 are categorized as adversarial, accidental, or structural (Ross, 2012). This research considers adversarial/malicious and accidental/error threat sources. The structure category is not included as threat sources here because it does not fit with the definition of a threat source that is defined in this research. Furthermore, in reviewing the subcategories provided in NIST SP 800-30, they represent vulnerabilities in the context of medical devices and are considered in that step of the assessment. For example, aging software or operating systems are vulnerabilities in our framework.

The FAIR methodology references the work of Intel (Rosenquist, 2009) in the establishment of a threat agent library. **Table 2** shows the attributes that we identify for establishing a threat agent library based on the work presented in the FAIR methodology (Freund and Jones, 2014) along with example threat actors and values. Each

of these characteristics bears significance in understanding the threat posed by each source. In developing the threat agent library, a panel is formed of security experts from within the organization and/or industry experts consulted with from outside of the organization who are familiar with the organization's security infrastructure and relevant threat intelligence. The threat agent library should be updated on a regular basis to reflect the state of threats to the organization.

Table 2: Threat Agent Library

Threat Source	Motive	Primary Intent	Sponsorship	Preferred Target characteristics	Preferred targets	Capability	Personal Risk Tolerance	Concern for Collateral Damage
Established Cyber-criminal Organization	Financial or PHI	Data gathering and/or disruption of services	unknown	Easy financial gains via remote means	Entities with financial resources or high value assets or IP	Well-funded trained and skilled	Very high	Medium
User (error)	Unmotivated Error	goodwill	none	Systems they access	no preference	Low to high depending on system	Low to medium	High

3.2.4 *Threat Source Capability Estimation*

For each threat source identified in the threat agent library, the capability of the source to compromise assets is estimated by the panel of security experts through an elicitation process. Experts should first be calibrated using established methods for improving the accuracy of expert judgment (Lichtenstein and Fischhoff, 1980). The experts in this elicitation would be security experts who have knowledge of the threat agents and the security infrastructure of the organization. The predictions elicited are a general assessment of each threat source's exploitation capability. Included in the assessment is the estimated threat actor capability - minimum, maximum, and most likely - and the expert's level of confidence in the estimate. We use a scale of 1 to 100 to for each estimate. These estimates will be used to perform a PERT distribution of the capability of the actor in measuring the threat event frequency.

Table 3: Threat Source Capability

Threat Source/Agent	Threat Type	Actor Threat Capability Min	Actor Threat Capability Max	Actor Threat Capability ML	Threat Actor Capability Confidence
Insider	Malicious	40	85	50	90
Cybercriminal organization	Malicious	60	90	80	90
User (error)	Error	40	90	50	90

3.2.5 *Identify Threat Events*

Threats events are specified as single events, actions, or circumstances. Threat events are characterized by the tactics, techniques, and procedures (TTPs) utilized by the threat source.

Threat events are identified by reviewing available documentation. For example, there may be documentation that identifies general threats discovered through regulatory compliance processes in the organization. Next, OSINT techniques and review of industry resources should be conducted to identify threat events. This review corresponds to the subcategory "ID.RA-2: Cyber threat intelligence is received from information sharing forums and sources" in the Cybersecurity Framework (NIST - CF, 2018). Additional threat events can be obtained from examples provided in NIST SP 800-30 (Ross, 2012). Threat event information may also be gathered from descriptions provided in CVEs discovered in the vulnerability identification process.

3.2.6 Develop Risk Scenarios

We develop risk scenarios for every combination of asset, target medical device, threat source, and CIA impact combination. We begin developing scenarios by listing all the vulnerabilities present for each medical device. For each of these pairs, we identify which assets could be impacted by the vulnerability. For each of these triplets, we identify the threat sources in the threat actor library that would likely be able to and/or interested in exploiting the vulnerability. Lastly, we identify each of the CIA effects that could occur for each of these possibilities. The resulting list are the risk scenarios that could result in exploitation of any of the vulnerabilities identified.

This list of risk scenarios should contain the following single-valued attributes: asset, target medical device, vulnerability, threat source, and potential CIA effect. **Table 4** shows a list of example scenarios.

Table 4: Example Risk Scenarios

Scenario	Asset	Target Device	Vulnerability	Threat Source	CIA Effect
AA	Asset 1	Medical Dev 1	CVE ID (or none)	Threat Source 1	C or I or A
AB	Asset 1	Medical Dev 1	CVE ID (or none)	Threat Source 2	C or I or A
AC	Asset 2	Medical Dev 2	CVE ID (or none)	Threat Source 1	C or I or A
..	..				
ZZ	Asset n	Medical Dev n	CVE ID (or none)	Threat Source n	C or I or A

In addition to the risk scenarios resulting from known vulnerabilities, consideration should be given for the possibility of unknown vulnerabilities being exploited. An organization may choose to take the approach that all networked medical devices could be exploited, or they may choose to evaluate them based on the characteristics of the device that were collected in the identification step. For any medical devices with no known vulnerabilities that could have a particular CIA effect, a group of risk scenarios should be developed considering the assets, the threat actors, and the CIA impacts.

3.2.7 Assign Threat Source Capability Estimate to Each Risk Scenario

For each of the risk scenarios, assign an estimate of the threat actor's capability to exploit the vulnerability resulting in the CIA effect, including the minimum, maximum, most likely, and confidence in the estimate. In most cases these values can be taken directly from the threat source capability estimate made earlier in the process. However, if the actor's capability estimate would be different for the given scenario, these values can be changed to reflect the expert opinion of the actor's capability in that scenario.

3.2.8 Assign Threat Event Frequency Estimate to Each Risk Scenario

For each risk scenario, employ expert elicitation to estimate the frequency with which they predict that a threat agent will act in a manner that could result in loss within the given time frame. As in the estimation of threat actor capability, experts not already calibrated should first be calibrated using established methods for improving accuracy (Lichtenstein and Fischhoff, 1980). These experts would be security experts who have knowledge of the security infrastructure of the organization and have reviewed available threat intelligence and vulnerability information. The timeframe is that which is identified by the organization in the preparation step. An example would be a one-year timeframe. The estimates include a minimum frequency, maximum frequency, a most likely frequency, and a confidence in the estimate. These values are used to calculate a PERT distribution of threat event frequency (2).

$$Threat\ Event\ Frequency\ (TEF) = \frac{Min\ TEF + (4 * Most\ Likely\ TEF) + Max\ (TEF}{6} \quad (2)$$

In addition to the expert elicitation, in cases where there is an identified CVE in the risk scenario, we find additional information to help us understand severity of the vulnerability and therefore, the frequency with which a risk scenario may be exploited. We use the exploitability subscore (NIST- NVD, n.d.) of the CVSS score (FIRST, 2019), discussed in the vulnerability identification step, to further refine the threat event frequency.

The exploitability subscore is derived from a combination of the CVSS metrics of attack vector (network, local, physical), level of attack complexity (high, low), privileges required (none, low, high), and whether interaction

with a user is required. The CVSS exploitability subscore is a numeric value between 0.12 and 3.9 based upon the CVSS v3.1 formula (FIRST, 2019). We use the exploitability sub score by converting it to a percentage of the range 0.12 to 3.9 to serve as a relative indicator of the vulnerability's exploitability, and we multiply this value by the PERT distribution of our expert's estimation of threat actor capability (TAC) as shown in (3).

$$\text{Vulnerability} = \frac{(CVE\ Expoitablity\ sub\ score - 0.12)}{3.9 - 0.12} x \frac{Min\ TAC + (4\ x\ Most\ Likely\ TAC) + Max\ TAC}{6} \quad (3)$$

The calculation for the loss event frequency is shown in (4).

*Loss Event Frequency = Threat event frequency * vulnerability* (4)

3.2.9 Assign Single Loss Event Magnitude Estimate to Each Risk Scenario

For each risk scenario, we employ expert elicitation to estimate the magnitude of a single loss event. This estimate is the magnitude of a potential single event loss without consideration for frequency. The experts in this elicitation would be business experts from within the organization. It may include financial experts, legal experts, or others who would have knowledge of the value of the assets and the potential impacts that could result from exploitation of a particular asset. The estimates include minimum loss, maximum loss, a most likely loss, and a confidence in the estimate. These values are used to calculate a PERT distribution estimate the single loss event magnitude. It will be multiplied by the frequency estimate gathered above to determine the magnitude within the timeframe chosen for the risk assessment.

In addition to the estimation of loss magnitude, in risk scenarios where there is an identified CVE, a loss magnitude multiplier is calculated using attributes of the CVE. We use the CIA impact (NIST- NVD, n.d.) of the CVSS score (FIRST, 2019), discussed in the vulnerability identification step, to further refine the loss magnitude. The CIA impacts each have a value of high, low, or none.

We choose a loss magnitude multiplier of 1.05 if the impact metric value is high and a multiplier of 1 for values of low or none. Our rationale is that the expert estimation is sufficient for anything that is not characterized as a high impact. Each risk scenario contains only one CIA impact, so there is only one multiplier for each scenario.

Table 5 shows the loss magnitude multipliers. The calculation for Single Loss Magnitude is show in (5).

Table 5: Loss Magnitude Multiplier

CVE CIA Impact value	Loss Magnitude Multiplier
If CVE CIA Impact = Low	1
If CVE CIA Impact = High	1.05
If CVE CIA Impact = None	1
If no CVE or known exploit	1

$$\text{Single Loss Magnitude} = Loss\ Magnitude\ Multiplier\ x\ \frac{Min\ Single\ Loss\ Mag\ Est. + (4 * Most\ Likely\ Single\ Loss\ Mag\ Est.) + Max\ Single\ Loss\ Mag\ Est.}{6} \quad (5)$$

3.2.10 Calculate the Expected Loss Magnitude for each Risk Scenario

The expected loss magnitude, or the quantified risk, is calculated for each risk scenario in consideration of the threat source, the threat source capability, the exploitability of the vulnerability, the impact effect, and the single loss expectancy, multiplied by the threat event frequency estimate to arrive at an estimate of the loss that could be experienced within the timeframe of the risk assessment. The calculation is shown in (6).

Risk = Expected Loss Magnitude = Single Loss Magnitude * Loss Event Frequency (6)

4 Discussion & Future Work

The results of the assessment can be communicated through reporting, sorting, and summarizing the details in a manner that is informative to the organization. In addition, Monte Carlo simulations can be performed using methods such as beta distributions.

The framework provides a method for quantifying risk, which to our knowledge is novel in the context of medical devices. The use of expert judgment is necessary in making predictions in domains such as cybersecurity that lack historical data or regularity. Methods for calibrating experts have been established and shown to improve expert judgment. We propose that a reduction in the uncertainty about the riskiness of the cybersecurity status of medical devices can be achieved using this framework.

The next step in the risk management process is to identify mitigations or controls that can reduce the loss magnitude. This is a combination of controls that are already in place and controls that can be implemented. In doing so, a process for quantifying the reduction in loss magnitude that may be done following the same strategy as has been used to estimate the loss magnitude and the threat actor capability here.

Automating the methodology used in this framework using a relational database with automation of inputs of vulnerability information can be conducted to make this scalable to the full medical device inventory of a healthcare organization.

References

Caralli, R., Stevens, J., Young, L. and Wilson, W. (2007), *Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process*, No. CMU/SEI-2007-TR-012, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

CIS. (2019), "CIS Critical Security Controls v7.1", Center for Internet Security, April.

Collier, K. (n.d.). "Baby died because of ransomware attack on hospital, suit says", *NBC News*, available at: https://www.nbcnews.com/news/baby-died-due-ransomware-attack-hospital-suit-claims-rcna2465 (accessed 29 June 2022).

Cooke, R.M. and Goossens, L. (n.d.). "Procedures guide for structured expert judgment", *EUR(Luxembourg)*, Office for official publications of the European Communities.

Eddy, M. and Perlroth, N. (2020), "Cyber Attack Suspected in German Woman's Death", *The New York Times*, 18 September.

FDA. (2018), "Overview of Medical Device Classification and Reclassification", *US Food & Drug Administration*.

Federal Bureau of Investigation (FBI). (2022), "220912 Unpatched and Outdated Medical Devices Provide Cyber Attack Opportunities", Federal Bureau of Investigation (FBI), 12 September.

FIRST. (2019), "Common Vulnerability Scoring System v3.1", FIRST, June.

Freund, J. and Jones, J. (2014), *Measuring and Managing Information Risk: A FAIR Approach*, Butterworth-Heinemann.

Ganin, A.A., Quach, P., Panwar, M., Collier, Z.A., Keisler, J.M., Marchese, D. and Linkov, I. (2020), "Multicriteria decision framework for cybersecurity risk assessment and management", *Risk Analysis*, Wiley Online Library, Vol. 40 No. 1, pp. 183–199.

Goossens, L.H., Cooke, R., Hale, A.R. and Rodić-Wiersma, L. (2008), "Fifteen years of expert judgement at TUDelft", *Safety Science*, Elsevier, Vol. 46 No. 2, pp. 234–244.

Hassan, N.A. and Hijazi, R. (2018), *Open Source Intelligence Methods and Tools*, Springer.

HIMSS. (2018), *2018 HIMSS Cybersecurity Survey*, HIMSS North America, Chicago, IL.

HIMSS. (2022), *2021 HIMSS Healthcare Cybersecurity Survey Report | HIMSS*.

Hireche, R., Mansouri, H. and Pathan, A.-S.K. (2022), "Security and Privacy Management in Internet of Medical Things (IoMT): A Synthesis", *Journal of Cybersecurity and Privacy*, MDPI, Vol. 2 No. 3, pp. 640–661.

HITRUST Alliance. (2019), "HITRUST CSF, Version 9.2", HITRUST Alliance, January.

Hubbard, D.W. and Seiersen, R. (2016), *How to Measure Anything in Cybersecurity Risk*, John Wiley & Sons, Inc., Hoboken, NJ.

ISO. (n.d.). "ISO - Standards", *ISO*, available at: https://www.iso.org/standards.html (accessed 11 October 2022).

Kaplan, S. and Garrick, B.J. (1981), "On The Quantitative Definition of Risk", *Risk Analysis*, Vol. 1 No. 1, pp. 11–27, doi: 10.1111/j.1539-6924.1981.tb01350.x.

Krisper, M., Dobaj, J. and Macher, G. (2020), "Assessing Risk Estimations for Cyber-Security Using Expert Judgment", in Yilmaz, M., Niemann, J., Clarke, P. and Messnarz, R. (Eds.), *Systems, Software and Services Process Improvement*, Vol. 1251, Springer International Publishing, Cham, pp. 120–134, doi: 10.1007/978-3-030-56441-4_9.

Landoll, D. (2011), *The Security Risk Assessment Handbook: A Complete Guide for Performing Security Risk Assessments*, 2nd ed., CRC Press, Inc., Boca Raton, FL, USA.

Lee, I., Sokolsky, O., Chen, S., Hatcliff, J., Jee, E., Kim, B., King, A., *et al.* (2012), "Challenges and Research Directions in Medical Cyber Physical Systems", *Proceedings of the IEEE*, Vol. 100 No. 1, pp. 75–90, doi: 10.1109/JPROC.2011.2165270.

Lichtenstein, S. and Fischhoff, B. (1980), "Training for calibration", *Organizational Behavior and Human Performance*, Elsevier, Vol. 26 No. 2, pp. 149–171.

Martin, T.G., Burgman, M.A., Fidler, F., Kuhnert, P.M., Low-Choy, S., McBride, M. and Mengersen, K. (2012), "Eliciting expert knowledge in conservation science", *Conservation Biology*, Wiley Online Library, Vol. 26 No. 1, pp. 29–38.

NIST. (n.d.). "Predisposing Condition - Glossary | CSRC", *NIST Computer Security Resource Center*, available at: https://csrc.nist.gov/glossary/term/predisposing_condition (accessed 17 July 2022).

NIST - CF. (2018), "Framework for Improving Critical Infrastructure Cybersecurity Version 1.1", National Institute of Standards and Technology, 16 April.

NIST - CPE. (n.d.). "Common Platform Enumeration", *National Institute of Standards and Technology (NIST)*, Government, , available at: https://csrc.nist.gov/Projects/Security-Content-Automation-Protocol/Specifications/cpe (accessed 4 March 2018).

NIST - SCAP. (n.d.). "Security Content Automation Protocol | CSRC", *NIST*, Government, , available at: https://csrc.nist.gov/projects/security-content-automation-protocol/ (accessed 4 March 2018).

NIST- NVD. (n.d.). "National Vulnerability Database", *Government*, available at: https://nvd.nist.gov/ (accessed 20 November 2017).

NTIA. (2019), *Software Component Transparency*, National Telecommunications and Information Administration, p. 20.

Ortiz, N.R., Wheeler, T.A., Breeding, R.J., Hora, S., Meyer, M.A. and Kenney, R.L. (1991), "Use of expert judgement in NUREG-1150", *Nuclear Engineering and Design*, Netherlands, Vol. 126 No. 3, pp. 313–331.

Reel, M. and Robertson, J. (2015), "It's Way Too Easy to Hack the Hospital", *Bloomberg.Com*, November, available at: http://www.bloomberg.com/features/2015-hospital-hack/ (accessed 17 April 2016).

Rosenquist, M. (2009), "Prioritizing information security risks with threat agent risk assessment", Intel Corporation White Paper.

Ross, R. (2012), *Guide for Conducting Risk Assessments - Special Publication (NIST SP) - 800-30 Rev 1*, NIST Special Publication No. 800–30 Rev 1, National Institute of Standards and Technology, Gaithersburg, MD.

Samonas, S. and Coss, D. (2014), "The CIA strikes back: Redefining confidentiality, integrity and availability in security.", *Journal of Information System Security*, Vol. 10 No. 3.

TrapX Labs. (2015), *MEDJACK (Medical Device Hijack)*, TrapX Security, Inc., p. 39.

US Dept of Homeland Security. (2019), "A Lifeline: Patient Safety & Cybersecurity", U. S. Department of Homeland Security.

Verizon. (2022), "2022 Data Breach Investigation Report (DBIR)", *Verizon Enterprise Solutions*, available at: https://www.verizon.com/business/resources/reports/2022/dbir/2022-dbir-data-breach-investigations-report.pdf (accessed 18 June 2022).

Williams, P.A. and Woodward, A.J. (2015), "Cybersecurity vulnerabilities in medical devices: a complex environment and multifaceted problem", *Medical Devices (Auckland, NZ)*, Dove Press, Vol. 8, p. 305.

Towards the Usefulness of Learning Factories in the Cybersecurity Domain

N Veerasamy, T Mkhwanazi and Z Dawood
CSIR, Pretoria, South Africa

nveerasamy@csir.co.za
tmkhwanazi@csir.co.za
zdawood@csir.co.za

Abstract: The success of an organisation depends on its employees' skills and the extent to which they are developed. Although organisations often assume employees are fit and ready for a new position or new developments in their functions, employees need adequate training before, during and after effective performance in their respective roles. Amongst other important roles, training is significant in problem-solving, continuously improving skills, and creating consistency or culture in the work environment. Nonetheless, the significance of training is often disregarded or not understood by organisations as there are often inadequacies, inconsistencies, and ignorance from the employer. Furthermore, organisations are facing cybersecurity skills shortages. Some specialists leave the profession due to a lack of skills or support. The lack of experienced and qualified cyber security specialists increases the risk of IT system systems being targeted with cyber-attacks. Having insufficient cybersecurity staff, companies may struggle to protect their networks from attacks. Organisations are being placed into a troubling position as the threat landscape continues to evolve. With the growth in volume and sophistication of cyber security attacks, the problem of a skilled workforce is exasperated. In order to support the cybersecurity workforce, this paper proposes the implementation of learning factories. Typically, learning factories have been used in the manufacturing sector. However, the fundamental principles and guiding ideologies can also be applied in the cybersecurity domain. Learning factories provide a mechanism to remove the barriers of entering the field of cybersecurity by cultivating and nurturing a cybersecurity workforce. They enable the broadening of the scope for talent and change our current working practices and tighten the gap between education and experience. The closing of the talent gap is an important imperative for cybersecurity. In this paper, a motivation and description of the functionality of learning factories for cybersecurity is provided. Through this paper the benefits of learning factories will be highlighted in order to show the advantages of active engagements in learning activities, real-world application and information sharing.

Keywords: Cybersecurity, Information Security, Learning factory, Experiential Learning, and Cybersecurity skills

1. Introduction

Research has demonstrated that there is a cybersecurity skills shortage globally. The Joburg Centre for Software Engineering (JCSE) annual surveys have reported an acute shortage of skilled information and communications technology (ICT) workers in South Africa, with information security a leading issue for employers (Schofield, 2016). Tertiary education institutions do not possess the necessary responsiveness. According to Pew Research Centre, mobile access penetration in South Africa as of 2018 was 95% (Pew, 2019); only 5% of adults do not have access to a cell phone. Due to the lack of cyber security skills, this has opened society to vulnerabilities and a marked increase in cybercrime. To deal with this shortage of skilled information security personnel, South Africa needs to develop a new pool of youth with cybersecurity skills.

The lack of cybersecurity skills is not an easy problem to solve. Typically, students need to enrol at a tertiary level institute, complete a degree or diploma, and thereafter supplement this with certifications. This process can take a few years in which organisations still actively look for skilled specialists,. To assist with this, the researchers investigate learning factories in the cybersecurity domain. There are various cybersecurity training and course offerings, but there is a disconnect between such training and the end-goal of developing skilled information security personnel. Most of the learning factories exist in the manufacturing domain, and there are a few cases of using learning factories for advancing manufacturing with the fourth industrial revolution (Baena et. al, 2017, Elbestawi et. al, 2018, Faller and Feldmüller,2015).

Integrating such training and course offerings into a learning factory can assist with the problem at hand. The research objective that this work intends is to investigate attributes for a cybersecurity learning factory. This will feed into a bigger objective, i.e., whether a cybersecurity learning factory can deal with the shortage of skilled information security personnel.

Learning factories can provides a useful mechanism to up-skill and train cyber security specialists. Due to the global cybersecurity shortage, it is imperative to propose innovative methods to empower and maintain cyber security skills. This paper will provide insight into how learning factories can be applied, as well as their benefits.

The structure of the paper is as follows: Section 2 provides a description of the usefulness of learning factories by initially looking at the manufacturing sector and then the application of learning factories to cyber security. Section 3 discusses the process of experiential learning in learning factories. Section 4 provides an overview of the contextual attributes of learning factories.

2. Usefulness of Learning Factories

2.1 Manufacturing Sector

Learning factories initially emerged as a teaching and learning concept out of the manufacturing sector, to train workers on how to apply Lean Manufacturing methods, among others. According to Monetti et. al (2022), Lean Manufacturing (LM) methods are typically applied by factory workers in the production line for three main objectives, namely:

- To foster skills to workers;
- To identify and eliminate non-value adding steps and processes in the production line; and
- To continually improve worker and manufacturing performance.

One of the well-known LM methods is the Kanban system. Kanban is a Japanese word meaning 'signboard'. The Kanban system means using visual cues to prompt the action needed to keep the process flowing. Kanban is a scheduling system developed to improve the efficiency of a manufacturing system. It was developed by a Toyota engineer who observed how retailers stocked their shelves- they would only order supplies as the old stock was dwindling. Therefore, Kanban is also known as just-in-time manufacturing which is to maintain efficiency in industrial production (Powell, 2018). There are several other LM methods such as Value Stream Mapping (VSM), Kanban/Pull production, 5S, and production smoothing" (Monetti et al., 2022). Workers require practical training to appropriately implement the latter tools. Learning factories were originally used to facilitate practical training for factory workers to implement LM in an environment that resembles a real factory assembly line. By definition, a learning factory is an integrative educational environment, designed to promote learning by training individuals to practically apply concepts, tools and cognitive skills to solve problems in an environment that is set up to resemble a real working environment (Enke et al., 2016, Mladineo et al., 2016, Monetti et al., 2022). Learning is often fostered when individuals are given the opportunity to put into practice their skills and not being shown something or taught in the classroom.

The focus of modern learning factories in the manufacturing sector is to advance employees' skills, agility and trust as demands in the work environment are consistently rising, compelling flexibility and strategy in problem solving across the board (Barton & Delbridge, 2001). The value of a learning factory in modern manufacturing is to upskill vertically, horizontally and across the organisational hierarchical (Tisch et al., 2015). The latter view is derived from developed manufacturers, such as Japan where the Kanban LM method originates, which place human resources, productive and production processes at the center of industrial success (Barton & Delbridge, 2001). Organisations thrive when employees' skills and development track record is strong. Therefore, learning factories offer practical and complex employee training which is an investment that can improve growth, application areas and competencies according to domains specifics (Tisch et al., 2015). The following paragraphs expand on the latter discussion.

Organisations may assume that employees are skilled and ready for a new position or new requirements in their roles, however employees need suitable training to ensure effective performance (Barton & Delbridge, 2001). Amongst other important roles, training is significant in problem-solving, continuously improving skills, creating consistency or culture in the work environment. However, the value of training is often disregarded or mis-understood by organisations, resulting in inadequacies, inconsistencies, and unawareness from the employers' side. Some of the inadequacies of training found by researchers are as follows (Barton & Delbridge, 2001):

- The effectiveness of worker training is not measured;
- The nature of training and training is not appropriately considered;
- Training initiation is not standardised; minimal evidence of training following promotions; and
- There is no shared or outlined concept regarding the purpose of training

Production and productivity requirements are changing in the manufacturing Sector. There are greater demands on workers as there is an increasing need for continuous innovation (Barton & Delbridge, 2001). The latter authors conclude that there is a greater need for active worker involvement and a transferral of decision-making roles that were previously reserved for managers to lower position workers. Furthermore, the role of training for supervisors and managers is shifting to be more focused on interpersonal skills development. The challenge however is that there is a lack of appropriate training facilities. Consequently, there is a lack of continuity and consistency in facilitated training which results in workers not having an appreciation of the significance of training (Barton & Delbridge, 2001). A practical example on the suitability of a training facility is that production plants and roles in an automotive factory are different, other workers may require technical skills while others require assembly or administrative skills. Therefore, it is important to design training offerings with envisaged skillset, environment and nature of work in mind.

Learning factories are commendable in resolving the issues discussed above. Learning factories are described as efficient systems in facilitating strategic and complex training processes that can go beyond competency building, to innovation as well (Zancul et al., 2022). Learning factories differ from traditional learning in that they equip individuals with applicable knowledge and principles to practise the lessons that were offered in a correct and intended manner. Learning factories provide an additional layer of learning in the form of "hands-on experiential learning" which trains participants through practical and direct engagement with the knowledge being imparted (Monetti et al., 2022, p. 1). Furthermore, learning factories foster experiential learning through simulation, an immersive teaching and learning method that has proved to be effective in captivating, engaging and intriguing participants as they learn (Monetti et al., 2022). Simulation is effective as participants are able to implement the concepts, tools or equipment provided in the learning factory, and observe its effects in real time and environment (Monetti et al., 2022). Subsequently and in accordance to "Bloom's taxonomy model in the cognitive domain" (Bloom, 1956), not only will the participants understand and remember the imparted knowledge— they will also be able to apply the knowledge appropriately, analyse its efficacy thus evaluating their competency in solving problems, while further improving personal and organisational performance by creating unique and innovative solutions. Learning factories, therefore, are the breeding ground for competency development and innovation.

Importantly, learning factories are not limited to the manufacturing sector, they can be adapted and applied in various sectors and environments that require practical skills (Enke et al., 2016). It has therefore become common practice for academic institutions to set up learning factories, especially for the built environment and engineering courses, in order to tackle real industry challenges (Monetti et al., 2022). The next section discusses the effectiveness of learning factories in fostering cybersecurity skills and competencies.

2.2 Cybersecurity Learning Factories

Learning factories may evolve out a collaboration of various stakeholders like corporate organisations, research institutions, and universities. The aim is to provide strategic guidance through the use of educational tools that help build competence, skills development as well as carry out research and development (R&D) and innovation.

Cybersecurity is conventionally perceived as solely technical, and subsequently tackled from a technical perspective only. However, cybersecurity is as political, social and economic as it is technical in nature. Therefore, to comprehend cybersecurity, a multi-dimensional approach and perspective is required. Researchers (Ramluckan et. al , 2019) note the theoretical acknowledgement of the need for a multi-disciplinary collaboration in cybersecurity research studies, and the offsetting lack of practicality in achieving it. Subsequently, the challenge of inadequate cybersecurity skills and innovation persist. To this extent Ramluckan et al. recommend the following (2019):

- Allowance for multi-dimensional research methodologies in cybersecurity studies;
- An ease in the "rigid educational structure" (i.e., NQF) of South African Higher Education Institutions (HEI); and
- Exclusive cybersecurity courses or qualifications, which are not offered in South African HEI, except as modules in technical or engineering courses, thus excluding social or legal courses for example.

Learning Factories present an imperative opportunity to advance the ideal multi-dimensional cybersecurity skills and innovation in the field. This is due to the nature of the learning factory being collaborative and involving diverse stakeholders, such as government, private or public sectors, industry experts and academic institutions who can offer diverse contributions (Zancul et al., 2022). Furthermore, the flexibility of learning factories to be setup physically and or virtually (Enke et al., 2016) may be advantageous given the physical restrictions resulting

from the global Corona Virus pandemic and the persistent digital divide. Therefore, there is a need for conducive cybersecurity learning factories that will facilitate national and international collaborations and sharing of holistic and multi-dimensional skills that are required to potentially solve the global cybersecurity challenges the world is faced with.

2.3 Application Example

"Adversary Thinking" is a commendable skill to possess to participate in or achieve cybersecurity goals. Adversary thinking is the ability of a cybersecurity professional to act like an attacker in order to devise effective defence mechanisms (Švábenský et. al, 2018). Such a skill can be cemented through a hands-on approach, such as interactive educational games to measure and assess student development (Švábenský et. al, 2018). Learning factories, through simulation, enable a secure environment to train individuals for adversary thinking.

The increasing demand for cybersecurity experts in the labour market exceeds the supply as there is a persistent shortage of skilled personnel. Therefore, there is a need for work-readiness programs for undergraduate students in higher learning institutions. To this effect, a hands-on learning environment that realistically simulates threats and attacks in cyberspace may be developed to teach computer networks and information security students practical defense mechanisms (Švábenský et. al, 2018). Adapting cybersecurity education into a game is a well advocated teaching and learning method among other immersive methods. Therefore, practical cybersecurity skills such as penetration testing may be offered in the form of a game in a learning factory, which promotes learner participation and engagement. The use of games in a learning factory supports traditional teaching and learning methods (Švábenský et al., 2018).

The International Academy for Production Engineering outlined the key characteristics of a learning factory (Abele et. al, 2015). We have adapted this to the cybersecurity domain for the application example in Table 1. Characterisation of the cybersecurity domain is the first step towards conceptualisation.

Table 1: Application example mapping to key characteristics of a cybersecurity learning factory (Adapted from Abele et. al, 2015)

Dimension	Feature	Cybersecurity application
Purpose	teaching and/or training and/or research	Teaching and training with skills on penetration testing for infrastructure or web-based application using University students/interns to address the cybersecurity skills shortage in South Africa.
Process	authentic + multi stations + technical and organizational aspects	The process involves using cyber security simulation and training tools hosted on computers.
Setting	changeable + real or virtual value chain	The learning environment will hybrid. To initiate, it will be in-house instructor led and continuation can occur using online tools.
Product	physical or service	No products will be manufactured in this environment. Skills will be developed and challenges will be completed.
Didactics	concept based + formal and informal learning + own actions of trainees + on-site or remote learning	Pre and post assessments will be conducted. The initiation phase will be with the help of an instructor. Thereafter, participants will work on their own and be evaluated.
Operating Model	sustainable plan allows the ongoing operation	This project will be started using government funds and thereafter be supplemented from other initiatives.

3. Experiential Learning through Learning Factories

Before students or new graduates/employees can embark into the real workspace, learning factories could be used to gain practical knowledge and skills. In the cybersecurity field, particularly new employees can benefit from experiential learning to firstly imbibe knowledge and skills before entering the real-world domain. Previously internships, in-service training, student teaching and undergraduate exchange programs have been used for experiential learning. Learning factories also provide a suitable solution for exposure to diverse skillsets.

The Centre for Teaching and Learning considers experiential learning to contain the following elements (2022):

- Reflection, critical analysis and synthesis;
- Opportunities for students to take initiative, make decisions, and be accountable for the results;
- Opportunities for students to engage intellectually, creatively, emotionally, socially, or physically; and
- A designed learning experience that includes the possibility to learn from natural consequences, mistakes, and successes.

Kolbs proposed a cycle of learning in 1984 that also depicts experiential learning (see Figure 1). It entails (Kolb, 1984):

- Direct engagement with the authentic situation for concrete experience;
- Observing the information and relating it to past experience and conceptual understanding;
- Distillation of perceptions into abstract concepts; and
- Applying new ideas and practising skills in a new proficiency

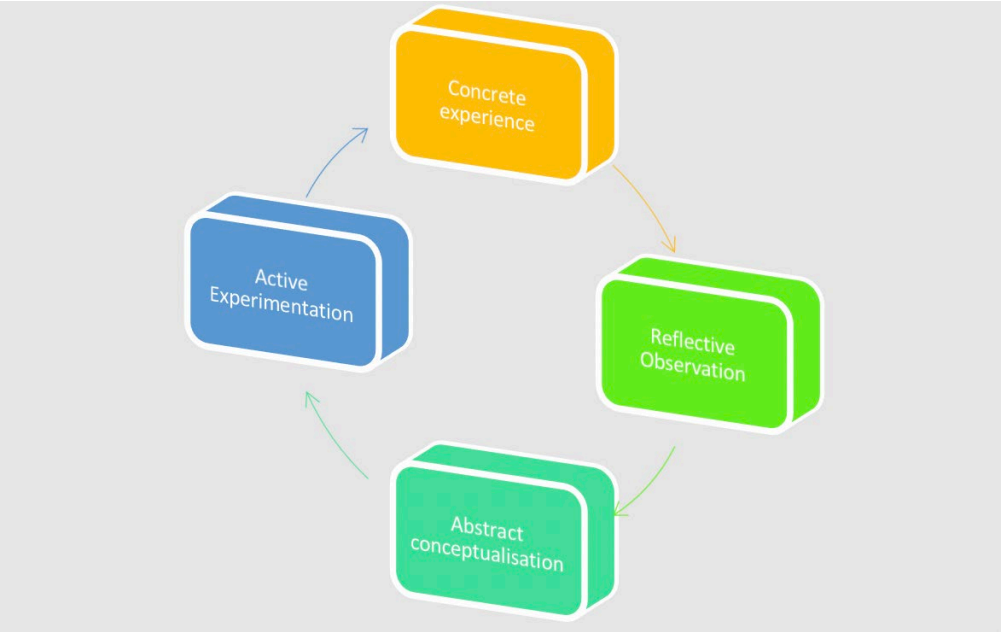


Figure 1: Cycle of Experiential Learning (Kolb, 1984)

Learning factories enable learners to investigate and experiment, be exposed to challenging situations and discover new solutions. It provides an ideal platform for learner to identify connections between concepts as well as apply theory and practical knowledge.

Learning factories serve three purposes (Monetti et. al, 2022). Firstly, recreating a real industrial or work environment either physically or through emulation, in the classroom. By classroom, this study is referring to any learning environment. Secondly, learning factories can recreate the conditions (e.g., setup and time constraints) in the industrial environment, in the classroom. Lastly, learning factory complement traditional teaching and learning processes by providing action-based, participatory, and interactive training. Learning factories support the concept of experiential learning by integrating the core processes of learning with experience offering in a suitable environment.

Conventionally, learners are taught the theoretical concepts of a particular field, however, they are not trained on when and how to implement those principles in the real world. Experiential learning is somehow transferred to the workplace, which results in incompetency. However, learning factories solve that problem in its entirety as they provide learners with hands-on experience, especially those in practical study fields or production focused fields such as manufacturing.

Facilitators are able to pose suitable problems to the students and provide support and resources to facilitate the learning process. Learning factories provide the opportunity for experiential leaning whereby students gain experience through the application of empirical training allowing them to investigate, explore, reflect and synthesise the knowledge and skills. Targeting learning activities can be planned. This can include hands-on experiments, simulations, training exercises, labs, practical exercises, and information sharing sessions. Figure 2 shows a representation of the experiential learning that is achieved through learning factories.

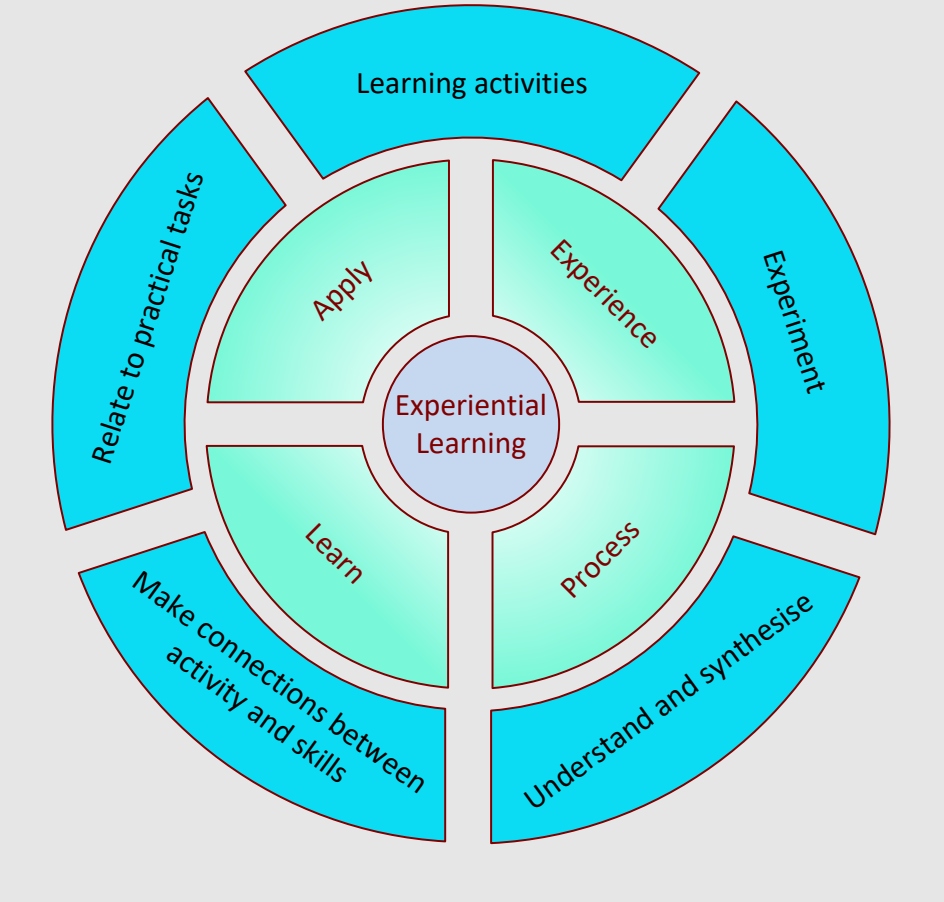


Figure 2: Experiential Learning

Experiences or learning activities can be selected based on their learning potential- whether it will enhance existing skills or pose a new situation/skill. During the experiential learning process in a learning factory, the learner will actively engage in processing information, understanding of concepts, investigating, constructing meaning, and overall being challenged to be creative or solution driven. After interacting in the domain of the learning activity/experience, knowledge will feed into the minds of the learner which then has to be processed. Through the initial observation and latter reflection, the learner can help understand the subject matter better, synthesis the knowledge and lead to more critical thinking and skills development. During this learning process, the learner will make connections between the different activities and skills. Later on, the learner can then apply the knowledge into other practical tasks.

4. Contextual Attributes of a Learning Factory

Learning factories facilitate action- and practice-based training and learning by compelling the learners to think and apply themselves rather than doing or acting only. As such learning factories are efficient in building competency and skills. Learning factories systematise and transfer domain-specific knowledge to the trainee. The implementation of learning factories is not limited and may therefore be adapted to suit custom requirements (Enke et. al, 2016).

Up until now, learning factories have not been typically used for the cybersecurity field. To initiate this application, the researchers have designed novel contextual attributes. Thus, with these contextual attributes, the usefulness of learning factories can be seen in Figure 3. The contextual attributes demonstrate the usefulness and application areas of Learning Factories. It aims to indicate what purpose learning factories have, how they can be operated, and the value gained from their use.



Figure 3: Contextual Attributes of Learning Factories

Figure 3 proposes five contextual attributes of learning factories, below is a brief description of each attribute.

- **Function** indicates the purpose and usefulness of learning factories such as training, skills development, competencies, and research. Enke, Tisch and Metternich refer to the purpose of training, education and research in their design dimension model of learning factories (2016).
- **Mode** refers to the planned objectives of learning factories in terms of achieving strategic partnerships, operational functionality or exchange mechanisms and information sharing
- **Benefits** includes the advantages of co-operation, collaboration, transformation and establishing a skilled workforce
- **Methodology** is the operating model of a cyber security learning factory that will provide for an agile, adaptive, active, responsive and engaging methodology. This is due to the vast number of activities that can be applied into a cyber security learning factory (scenarios, training, software, games, quizzes, simulations, practical exercises, tests, and presentations, videos)
- **Features** due to the remote capability of Information and Communication Technology, a cyber security learning factory has the advantageous features of allowing remote access, offering hybrid solutions and extendible capabilities (adapting or adding more labs, exercises, software, simulations, etc). Extendible capabilities are key in a fast-paced cyber security environment when the threat landscape and technologies are constantly evolving.

5. Conclusion

The global shortage of cyber security skills has cascaded to a national level. While most South African organisations at a private or public level have adopted using technological solutions to solve socio-economic problems, they do not possess the cyber security skills that are so relevant to ensure safety and security. Learning factories have been proposed in the manufacturing sector to offer a low-cost solution to replicating manufacturing environments and assisting in skills development. The authors provide adopting learning factories for the cyber security domain. To demonstrate the benefits of learning factories for the cyber security domain, a contextual attribute model was designed. This multi-dimension model highlights the qualities for a cyber security learning factory and aims to serve as the foundation for the development of cyber security learning factories. Future work will include delving deeper on methodologies for the design and execution of learning factories in the cyber domain.

References

Abele, E., Metternich, J., Tisch, M., Chryssolouris, G., Sihn, W., ElMaraghy, H., ... & Ranz, F. (2015). Learning factories for research, education, and training. *Procedia CiRp*, *32*, 1-6.

Baena, F., Guarin, A., Mora, J., Sauza, J., & Retat, S. (2017). Learning factory: The path to industry 4.0. Procedia manufacturing, 9, 73-80.

Barton, H & Delbridge, R. (2001). Development in the learning factory: Training human capital. Journal of European Industrial Training. 25. 465-472. 10.1108/03090590110410313.

Bloom, B. S. (1956). Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives: the classification of educational goals. Handbook, 1.

Centre for Teaching & Learning, 2022, Experiential Learning, Available at https://www.bu.edu/ctl/guides/experiential-learning/, Accessed 7 September 2022.

Elbestawi, M., Centea, D., Singh, I., & Wanyama, T. (2018). SEPT learning factory for industry 4.0 education and applied research. Procedia manufacturing, 23, 249-254.

Enke, J & Tisch, M & Metternich, Jo. (2016). Learning Factory Requirements Analysis – Requirements of Learning Factory Stakeholders on Learning Factories. Procedia CIRP. 55. 224-229. 10.1016/j.procir.2016.07.026.

Faller, C., & Feldmüller, D. (2015). Industry 4.0 learning factory for regional SMEs. Procedia Cirp, 32, 88-91.

Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. Englewood Cliffs, NJ: Prentice-Hall.

Mladineo, M., Gjeldum, N., & Veza, I. (2016). Lifelong Learning in Learning Factory. 23rd EurOMA Conference, Trondheim, Norway. https://www.researchgate.net/publication/349553755_Lifelong_Learning_in_Learning_Factory

Monetti, F. M., Boffa, E., Maffei, A., & de Giorgio, A. (2022). The Impact of Learning Factories on Teaching Lean Principles in an Assembly Environment. FAIM 2022: Flexible Automation and Intelligent Manufacturing International Conference, Detroit, MI, United States.

Powell, D. J. (2018). Kanban for lean production in high mix, low volume environments. IFAC-PapersOnLine, 51(11), 140-143.

Ramluckan, T., van Niekerk, B., & Leenen, L. (2019). Research Challenges for Cybersecurity and Cyberwarfare: A South African Perspective. 18th European Conference on Cyber Warfare and Security, Coimbra, Portugal. https://www.researchgate.net/publication/334327321_Research_Challenges_for_Cybersecurity_and_Cyberwarfare_A_South_African_Perspective.

Schofield A & Dwolatzky B. (2021). JCSE-IITPSA ICT Skills Survey, Available at https://www.iitpsa.org.za/wp-content/uploads/2021/09/2021-JCSE-IITPSA-ICT-Skills-Survey.pdf , Accessed 9 September 2022.

Švábenský, V., Cermak, M., Vykopal, J., & Laštovička, M. (2018). Enhancing cybersecurity skills by creating serious games. ITiCSE 2018: 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, New York, NY, USA. https://doi.org/10.1145/3197091.3197123.

Naïve Bayes Supervised Learning based Physical Layer Authentication: Anti-Spoofing techniques for Industrial Radio Systems

Andreas Weinand[1], Christoph Lipps[2], Michael Karrenbauer[1] and Hans Dieter Schotten[1,2]
[1]TU Kaiserslautern, Institute for Wireless Communication and Navigation, Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence, Intelligent Network Research Group, Kaiserslautern, Germany

weinand@eit.uni-kl.de
Christoph.Lipps@dfki.de
karrenbauer@eit.uni-kl.de
schotten@eit.uni-kl.de, Hans_Dieter.Schotten@dfki.de

Abstract: Physical Layer Security (PLS) based authentication schemes are an alternative to conventional security schemes such as e.g. certificates or Message Authentication Codes (MACs). They can provide a more lightweight solution compared to traditional cryptography in order to meet the requirement of secure data transmission. However, errors in Physical Layer Authentication (PLA) techniques can occur due to adverse influences resulting from PLS schemes such as receiver noise or fading channels. Skillfull methods are therefore required in order to detect anormal system behaviour. One promising solution are supervised classification schemes. The application of Naïve Bayes (NB) classifiers for PLA is therefore proposed and evaluated within this work. Prior to that, we analyse the resource efficiency within typical Ultra Reliable Low Latency Communication (URLLC) applications and conduct a security overhead analysis. We propose strategies in order to overcome the problem of missing training data from either the URLLC user or attacker node. A real world Software Defined Radio (SDR) based testbed using OFDM (Orthogonal Frequency Division Multiplexing) is implemented in order to evaluate the performance of NB based PLA. The measurements are conducted within the german campus network frequency band (LTE band 43). Further, we conduct a hyperparameter optimization (HPO) based on random search. The investigated classifiers show promising results in terms of authentication accuracy and Receiver Operating Characteristic (ROC) curve performance.

Keywords: Physical Layer Authentication; Physical Layer Security; Machine Learning; Ultra Reliable Low Latency Communication; Machine Type Communication

1. Introduction

Many Ultra Reliable Low Latency Communication (URLLC) services, such as automated driving systems, telesurgery, factory automation and many more are currently getting enabled by technologies such as 5G-New Radio (NR). Therefore, new approaches regarding system reliability, availability and security need to be considered in order to meet the high requirements for this type of services. Conventional solutions have limitations with respect to system efficiency, since they were originally designed for large packet internet traffic such as Mobile Broadband (MBB). This is especially the case for solutions with respect to information security. In URLLC scenarios the traffic is however characterized by short and periodic messages which are transmitted at deterministic time slots (e.g. within the order of 10s of ms (Bockelmann, et al. 2017) (Yilmaz, et al. 2015)).

1.1 URLLC message security overhead ratio

In order to transmit cryptographically secure messages within an URLLC service, it is typically required to meet the security goals of message confidentiality, authentication and integrity. While confidentiality can be achieved by encryption of the message payload, authenticity and integrity are typically ensured by attaching authentication tags (e.g. Message Authentication Codes (MACs)) to messages. This leads however to an increase of the original payload length. For MACs the recommendation of the Internet Engineering Task Force (IETF) is to either use a cipher based MAC (CMAC), e.g. based on a Advanced Encryption Standard (AES)-128 block cipher, or a hash based MAC (HMAC) which is based on a cryptographic hash function. For AES-128 based CMAC a maximum shortening to 64 Bit is recommended (IETF, RFC 4493, 2006), while for HMAC a minimum MAC size of 80 Bit is recommended (IETF, RFC 2104, 1997). Additionally, the encryption process leads to a further increase of the message size due to zero-padding, which is typically not notable in broad band applications. For URLLC however this is not always the case if very short messages are used (e.g. within the order of 10s to 100s of Bits (Bockelmann, et al. 2017) (Yilmaz, et al. 2015)). Protection against replay attacks requires Further additional overhead (e.g. counters, time stamps). Figure 1 illustrates the overhead introduced by authentication tags, zero-padding and replay protection for different URLLC message sizes. The Security Overhead Ratio (SOR), which is

defined as the percentage of security overhead compared to the payload of a message depending on the URLLC payload of N Bit, is then calculated as

$$\mathrm{SOR} = \frac{ceil\left(\frac{N}{B}\right)B+L+K}{N} - 1,$$

where $ceil()$ is the ceiling function (maximum of the two nearest integer numbers), B the cipher block size (e.g. 128 Bit for AES-128 cipher) L the length of the MAC tag in Bit and K the length of the counter in Bit.

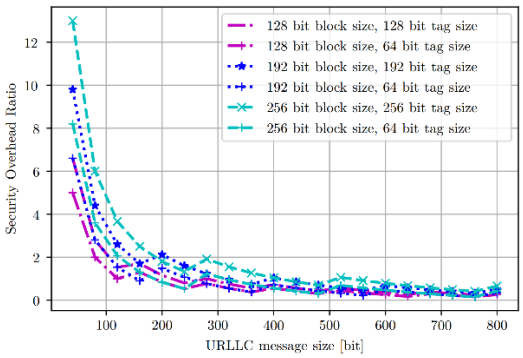


Figure 1: SOR for confidentiality, integrity and replay protection

The SOR is quite significant for short payloads and can reach up to 1000% of the URLLC message size. The SOR reaches 460% for a URLLC payload of 100 Bit, if a high level of security is applied (e.g. AES-256 cipher without MAC truncation and 48 Bit counter). The actual efficiency (application payload vs. overhead) on the air interface is even lower since this calculation is not considering the additional overheads introduced by non security functionalities. This work proposes therefore alternative solutions based on PLS for message authentication, integrity and replay protection in order to increase the ressource efficiency.

1.2 Organisation

The remainder of the work is organized as follows. In section 2 we give an overview of related work with respect to previously considered PLA approaches and in section 3 we introduce the system model. The NB classifiers which are used for Physical Layer Authentication (PLA) are presented within section 4. In section 5, we present techniques for artificial training data generation in case of the absence of an attacker during the training phase. Further, within sections 6 and 7 we present the evaluation methodology and results of the experimental evaluation respectively. Section 8 concludes the paper and provides an outlook for future work.

2. Related Work

Statistical methods were initially used for PLA, such as (Xiao, Greenstein, et al. 2007), where an approach based on simulation of the wireless channel and hypothesis testing is presented for static scenarios (no user mobility). It is later extended to mobile (time variant) scenarios in (Xiao, Greenstein, et al. 2008). The authors propose to use Neyman-Pearson Hypothesis Testing and least square adaptive filters in order to check if channel estimations of subsequent transmissions of a user are related to each other. The users they are considering have a low mobility (e.g. pedestrian speed) and they use an OFDM based transmission system for performance evaluation. The authors in (Tugnait und Kim 2010) propose a channel state information based authentication method for a single carrier system using whiteness of residuals testing. In (Baracca, Laurenti und Tomasin 2012), the authors consider a Multiple Input Multiple Output (MIMO) OFDM system and propose Generalized Likelihood Ratio Testing in order to detect attacks based on channel estimation data. They further introduce various attack strategies under which they evaluate the performance of their approach using Montecarlo simulations. They show how the misdetection probability is depending on different parameters such as the Signal-to-Noise Ratio (SNR), the number of subcarriers, the correlation coefficient, the probability of false alarms and knowledge of the attackers behaviour.

One of the first works that proposed machine learning (ML) based techniques in order to detect and identify transmitters based on physical layer metadata is (Gulati, et al. 2013), where the authors use Gaussian Mixture Models in order to detect an attacker. The feature they use are different antenna modes (angle of a leaky wave antenna) and they evaluate their proposed scheme within a static scenario at different parameter settings (number of different angles, attack probability, number of training messages, carrier frequency). Two approaches based on ML (Support Vector Machines and Linear Fisher Discriminant Analysis) are presented in

(Pei, et al. 2014). In (Forssell, et al. 2019), the authors use a single input single output system in order to detect the transmitter of data packets based on the respective line-of-sight components. The authors in (Weinand, et al. 2019) propose the application of supervised learning for channel based PLA within URLLC in a static environment. In (Wang, et al. 2017), the authors propose the usage of extreme learning machines for PLA. They further introduce a pseudo adversary model in order to generate artificial attacker training data. The authors in (Senigagliesi, Baldi und Gambi 2020) use Neural Networks and Support Vector Machines for PLA and evaluate the classifiers using a simulative framework. They addtionally propose to use one-class versions of the respective models in case of missing training data from the attacker node. Further, they conduct a model selection and manual HPO. They do however not consider cryptographic schemes in any phase of their proposed PLA procedure.

Though the results of most of the presented related works are promising, there are still several drawbacks existing. Many works use statistical decision metrics, which require a constant threshold to be set as input. This approach delivers typically only optimum performances in the cases which are evaluated. ML based classification schemes can therefore be applied in order to avoid that. Further, the outcome of simulative evaluations needs to be verified in real world scenarios with real hardware such as e.g. prototypes in order to prove their field performance. Cross validation should further be considered over a variety of classification schemes and parameterizations (e.g. HPO). Additionally, real world scenarios within the IR scope should be considered in order to pave the road towards standardization of PLA schemes within existing or future wireless systems. Channel measurements are therefore conducted within the 3.7 GHz campus network frequency band (Germany) within this work and realisitic URLLC traffic models are used in order to evaluate the PLA performance.

3. System Model

Within this section, we introduce the considered system model, as well as possible physical layer features to be used for PLA. Further, we introduce the considered attacker scenario.

3.1 Wireless System

We consider a wireless (public) channel between two users, Alice and Bob, intending to exchange authenticated messages with each other. A third user Eve, who is an adversarial user, tries to inject illegal messages masquerading as one of the legal users Alice or Bob. This can be done by reusing the radio ressources allocated to either Alice or Bob in case of centralized Radio Resource Management (RRM) or by spoofing their addresses in case of competetive/decentralized RRM. Figure 2 indicates the possible message flows over the wireless channel for all considered users. Messages transmitted by users are denoted with $\boldsymbol{x}_u$ and messages received by users with $\boldsymbol{y}_u$, where u denotes user Alice ($u = A$), Bob ($u = B$) or Eve ($u = E$).

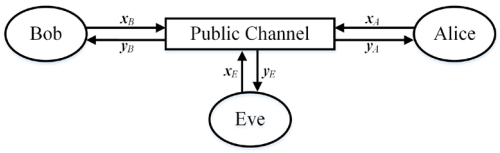


Figure 2: System Model

3.2 Physical Layer Metadata

Physical layer protocol metadata is used in PLS schemes in order to provide lightweight security functions. This metadata can either originate from the transceiver hardware (e.g. radio fingerprints or clock skew) or it can originate from characteristics of the wireless channel (e.g. RSSI or channel estimation data). In the first case it is already challenging for an attacker to spoof these characteristics by using default hardware (off-the-shelf). However, this requires an additional effort for the legitimate receiver in order to detect these characteristics and increases the receiver complexity and cost. For the second case, the respective data is typically available and accessible in any radio protocol. Therefore, we focus on channel based metadata and especially channel estimation data in our work, which is computed at the receiver side in form of a frequency domain channel estimation (OFDM based systems). The received signal at a user u_i in such systems is

$$\boldsymbol{y}_{u_i} = \boldsymbol{H}_{u_i u_j} \boldsymbol{x}_{u_i} + \boldsymbol{n}_{u_i u_j} \quad (1)$$

with

$$\boldsymbol{H}_{u_i u_j} = \left[h^0_{u_i u_j}, \dots, h^{M-1}_{u_i u_j} \right] \quad (2)$$

being the M-dimensional channel matrix of the frequency selective Single Input Single Output (SISO) fading channel between user u_i and u_j (transmitter) and $h^l_{u_i u_j}$ being the complex gain of the l-th sample in frequency domain ($l = 0, \dots, M-1$). It can be estimated as

$$\widehat{\boldsymbol{H}}_{u_i u_j} = \boldsymbol{H}_{u_i u_j} + \boldsymbol{\epsilon}_{u_i u_j} \quad (3)$$

by user u_i. Due to noise $n_{u_i u_j}$ which is modelled as a gaussian random variable with zero mean and variance $\boldsymbol{\sigma}^2_{n_{u_i u_j}}$ the channel estimation is not perfect and errors $\boldsymbol{\epsilon}_{u_i u_j}$ occur.

3.3 Attacker Model

A typical scenario for an attacker Eve is, that she is at a spatially different location compared to Bob and Alice (we assume a distance of more than the wavelength of the transmitted signal) and uses advanced equipment such as directed antennas and high sensitivity receivers in order to maximize the coverage to her benefits. We further assume perfect knowledge of the underlying communication protocol at Eve to run active attacks such as masquerade attacks, replay attacks or address spoofing attacks by introducing messages $\boldsymbol{x}_E$.

It is not assumed that Eve is gaining physical access to Alice or Bob to accomplish invasive attacks such as hardware modification. Further, other active attacks such as Denial-of-Service (DoS) attacks due to jamming are not considered. We assume that the legitimate communicating participants Bob and Alice have already carried out initial user authentication to each other and have set up trust in a secure way by using cryptographic certificates or MACs. Attacks on the initial authentication stage are not considered. Further, we assume that Eve is, depending on the current attack strategy, spoofing on transmission slots which are actually allocated to Bob by using a higher transmit power in order to make Alice overhear Bobs transmissions. In order to successfully carry out attacks, the receive power of Eves transmissions has to be higher than that of Bobs transmissions at Alice.

Within this work we consider that the attacker node is transmitting messages at a constant probability which is denoted Attack Intensity (AI). It gives the percentage of packets injected by Eve based on the total number of packets transmitted by Bob or Eve in total.

4. Naïve Bayes classification for PLA

After introducing the channel based PLA concept, we present the NB classifiers that were used in order to evaluated the PLA performance within our testbed.

4.1 Wireless channel based PLA

Within PLS based authentication such as PLA, receivers try to identify the transmitter based on e.g. channel characteristics instead of evaluating an authentication tag. Considering Bob transmitting messages $x_B(k)$ to Alice, Alice will estimate the channel as $\widehat{\boldsymbol{H}}_{AB}(k)$ when receiving message $\boldsymbol{y}_A(k)$ at times $k = 0, \dots, N_{train} - 1$. The messages $\boldsymbol{x}_B(k)$ sent by Bob need to contain cryptographic information from higher layers (above physical layer) in order to prohibit any spoofing attacks by Eve during time $k = 0, \dots, N_{train} - 1$. Alice authenticates Bob then using PLA from time $k \geq N_{train}$. This is realized by comparing the physical parameters such as channel state estimation or Channel Impulse Response (CIR) to the previously recorded versions at times $k = 0, \dots, N_{train} - 1$. Based on that, two hypothesis are possible about the origin of the respective estimated channel $\widehat{\boldsymbol{H}}(k), k \geq N_{train}$ and the payload of messages $y_A(k)$, which are

$$\mathcal{H}_0: \widehat{\boldsymbol{H}}(k) \text{ due to } \boldsymbol{x}_B(k) \text{ and}$$
$$\mathcal{H}_1: \widehat{\boldsymbol{H}}(k) \text{ due to } \boldsymbol{x}_E(k). \quad (4)$$

In case of $\mathcal{H}_0$ follows

$$\widehat{\boldsymbol{H}}(k) = \boldsymbol{H}_{AB}(k) + \boldsymbol{\epsilon}_{AB}(k) \quad (5)$$

and in case of $\mathcal{H}_1$

$$\widehat{\boldsymbol{H}}(k) = \boldsymbol{H}_{AE}(k) + \boldsymbol{\epsilon}_{AE}(k). \quad (6)$$

The reference channel estimation used for comparison needs to be updated due to temporal variations in wireless channels, such as due to doppler fading. Otherwise it is not possible to distinguish between channel variations and messages introduced by an attacker after some point. Therefore, the update interval should be below the channel coherence time T_C of the $\boldsymbol{H}_{AB}$ channel in order to catch up with these variations. The temporal difference between two subsequent channel estimations should further always be below T_C. Due to

such infuences on the physical parameters used within the PLA, errors can occur. Therefore, we propose the utilization of NB supervised classification schemes in order to distinguish between transmissions of users Bob and Eve received by Alice respectively.

4.2 NB based supervised classification

The classification algorithms, which are based on NB methods, are applying Bayes' theorem. Given the value of the class variable it assumes the conditional independence between every pair of features. Based on y as the target variable and the dependent feature vector $x = [x_1, \dots, x_n]$, Bayes' theorem states that

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)}. \quad (7)$$

From the naive independence assumption $P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y)$ follows for all i

$$P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\dots,x_n)}. \quad (8)$$

Due to the constant $P(x_1, \dots, x_n)$, the classification rule $P(y|x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$ can be derived, which further results in $y_{pred} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y)$. The terms $P(y)$ (frequency of class y in the training set) and $P(x_i|y)$ can be estimated using Maximum A Posteriori (MAP) estimation. The term $P(x_i|y)$ however can be estimated in different ways based on the assumptions which are made on its distribution (Zhang 2004). Therefore, different NB approaches exist which are introduced below and implemented in our testbed.

4.2.1 Gaussian Naïve Bayes classifier

The Gaussian Naïve Bayes (GNB) classification algorithm estimates $P(x_i|y)$ as

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\rho_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\rho_y^2}} \quad (9)$$

assuming the likelihood of the features is gaussian. In order to estimate ρ_y and μ_y, maximum likelihood estimation is used.

4.2.2 Complement Naïve Bayes classifier

The Complement Naïve Bayes (CNB) algorithm is especially suited for imbalanced data sets. It uses statistics, which are derived from the complement of each class in order to compute the weights w. These are calculated as

$$w_{\tilde{c}i} = \log \hat{\theta}_{\tilde{c}i} = \frac{w_{ci}}{\sum_j |w_{ci}|} \text{ with } \hat{\theta}_{\tilde{c}i} = \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha} \quad (10)$$

Where $N_{\tilde{c}}$ is the total number of samples within the opposite class and $N_{\tilde{c}i}$ are the repetitions of a sample in the opposite class (McCallum und Nigam 1998). Further, α_i is a smoothing hyperparameter and $\alpha = \sum_i \alpha_i$. Based on that, the CNB classification rule is

$$\hat{c} = \arg\min_c \sum_i x_i w_{ci} \quad . (11)$$

4.2.3 Multinomial Naïve Bayes classifier

The Multinomial Naïve Bayes (MNB) classifier assumes multinomially distributed data, where the distribution is parametrized by vectors $\theta_y = [\theta_{y1}, \dots, \theta_{yn}]$ for each class y. θ_{yi} is the probability $P(x_i|y)$ of feature i belonging to class y and estimated according to

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (12)$$

Here, $N_{yi} = \sum_{x \in T} x_i$ is the count of feature i appearing in a sample of class y within the training set T and $N_y = \sum_{i=1}^{n} N_{yi}$ is the total count of all features for class y. A smoothed maximum likelihood estimation of $\hat{\theta}_{yi}$ is achieved by introduction of the smoothening parameter $\alpha \geq 0$, which allows for either Laplace smoothing ($\alpha = 0$) or Lidstone smoothing ($\alpha < 1$) (Rennie, et al. 2003).

5 Artificial Training Data Generation for PLA

A big issue of the practical application of PLA is the fact, that the attacker is typically not present during the training phase. Different strategies can be deployed in order to overcome this, where artificial (attacker) training data is generated. We propose the utilization of fixed sequences for that purpose, since they are in contrast to the approach in (Wang, et al. 2017) independent of the channel profiles of the URLLC users and can therefore be generated in a preliminary manner reducing complexity and cost. The possible options we investigate here are 0-sequences

$$h_{AE}^{i} = 0, i = 0, \dots, M-1, (13)$$

1-sequences

$$h_{AE}^{i} = 1, i = 0, \dots, M-1, (14)$$

and random sequences

$$h_{AE}^{i} = \text{rand}(0,1), i = 0, \dots, M-1, (15)$$

where $\text{rand}(0,1)$ is a random float number within the range of 0 and 1. We are further assuming normalized channel profiles of the URLLC users as described in section 6.1.

6 Evaluation Methodology

The evaluation methodology used within section 7 is presented in this section including steps such as data pre-processing, hyperparameter optimization and PLA performance metrics.

6.2 Data Pre-processing

The pre-processing steps as depicted within Figure 3 are applied to the validation and testing datasets respectively, in order to fit the classifiers. The magnitudes $\boldsymbol{F}(k) = \left|\widehat{\boldsymbol{H}}(k)\right|, k = 0, \dots, N$ of the complex channel gains are calculated first for each data sample $\widehat{\boldsymbol{H}}(k)$. A pre-processing step for feature dimensionality reduction is added in order to investigate the impact of feature dimensionality. This reduces the number of features of the original sample from $M = 48$ to M_{red} by sampling the original feature vector at the respective rate yielding

$$\boldsymbol{F}_{red}(k, l^*) = \boldsymbol{F}\left(k, \frac{lM}{M_{red}}\right), \; l^* = 0, \dots, M_{red} - 1, (16)$$

where l^* denotes the index of the features in the new feature dimension M_{red} and l the index in the original feature dimension.

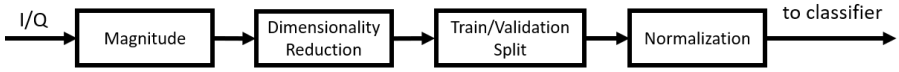


Figure 3: Data pre-processing

Afterwards, the data set is split into training and validation sets, which are used to fit the classifiers (training) and validate their performance. The training set is given by $\boldsymbol{F}_{train} = [\boldsymbol{F}_{red}(0), \dots, \boldsymbol{F}_{red}(N_{train} - 1)]$ and the remaining samples are used for validation. Finally, the training and validation data is normalized. This is done by calculating the mean and variance of the training set, then the training set is transformed to

$$\boldsymbol{F}_{train}^{norm} = \frac{\boldsymbol{F}_{train} - mean(\boldsymbol{F}_{train})}{std(\boldsymbol{F}_{train})} \; (17)$$

and the validation set is transformed according to

$$\boldsymbol{F}_{valid}^{norm} = \frac{\boldsymbol{F}_{valid} - mean(\boldsymbol{F}_{train})}{std(\boldsymbol{F}_{train})} \; . (18)$$

6.3 Hyperparameter Optimization

We use a random search in order to find the optimal hyperparameters

$$\lambda^* = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \underset{x_k^{valid}}{\operatorname{mean}} L(y_k^{train}, f_\lambda(x_k^{train})) \quad (19)$$

to the given classification task. Possible value ranges for each model and respective hyperparameter are given as input and the models are fitted and validated on all combination of hyperparameters. For this, we only use N_{valid} data sets of the overall $N_{data} = N_{valid} + N_{test}$ data sets. We use the accuracy metric in order to select the optimal hyperparameter set for each classifier. It is given by

$$P_A = \frac{1}{N} \sum_{k=N_{train}}^{N-1} \boldsymbol{y}_{pred}(k) - \boldsymbol{y}_{true}(k)\,, (20)$$

where $\boldsymbol{y}_{pred}$ is the label vector predicted by the respective classifier (user Bob or Eve is the transmitter) and $\boldsymbol{y}_{true}$ the true label vector. The hyperparameters derived from random search for the NB classifiers introduced in section 4.2 are listed in Table 1 accordingly. The remaining N_{test} data sets are used in order to evaluate their performance on the given classification task after deriving the optimal hyperparameters for each classifier. Beside the accuracy metric, ROC curves are used including Area Under Curve (AUC) scores for that purpose.

Table 1: Optimal hyperparameters for each classifier

Classifier	Parameter	Value
GNB	var_smoothing	$1.5189 \cdot 10^{-7}$
CNB	alpha	0.6213
MNB	alpha	0.5042

7 Performance Evaluation

In this section, we first describe the process of acquisition of channel estimation data. Then the evaluation scenario is presented as well as results and their discussion respectively.

7.2 Data Acquisition

We used Universal Software Radio Peripheral (USRP) B210 SDR platforms and an OFDM based transmission systems in order to obtain realistic channel estimation data. The OFDM system is implemented with GNURadio OFDM transmitter and receiver baseband processing blocks. These provide the functionality for channel estimation based on the Schmidl and Cox method (Schmidl und Cox 1997) using the respective preamble symbols. We use an OFDM setup with a Fast Fourier Transform (FFT) size of 64 with 48 active subcarriers. The cyclic prefix length is 16 samples, whereas the bandwidth is 10 MHz and the carrier frequency is set to 3.715 GHz. Each data packet consists of 2 preamble symbols, which are used for channel estimation and frequency offset estimation at the receiver and N data symbols. The data packets are transmitted with a periodicity of 10 ms, which yields us a vector of 48 complex symbols for each received OFDM packet (only the active subcarriers are considered for channel equalization).

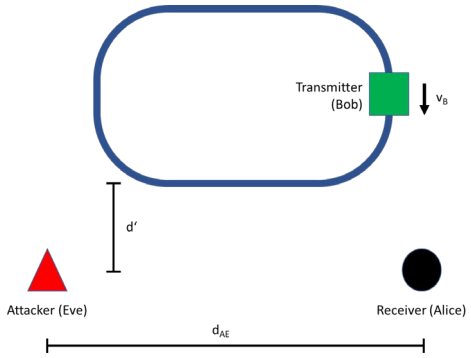


Figure 4: Experiment setup with positions of Alice, Bob and Eve

A mobile URLLC scenario is consider where user Bob is under mobility and emulating a URLLC user device transmitting messages, while Alice and Eve are static emulating the URLLC gateway receiving messages and the attacker node transmitting messages respectively (see **Figure 4**). The environment is a mixed office/lab area with a lot of objects and metal walls. Datasets of channel estimation data are recorded by Alice from N_{data} different position constellations of the users (Alice and Eve). In case of a transmission of Eve, it is assumed, as mentioned in section 3.3, that Bobs regular transmission is interfered by Eve (e.g. using a comparatively higher transmit power). The experiment parameters are summarized within Table 2.

Table 2: Validation and testing parameters

Parameter	Variable	Value
Carrier frequency	fc	3.715 GHz
Transmission frequency	ΔT	100 Hz
Channel bandwidth	BW	10 MHz
No. of datasets	N_{data}	10
No. of validation/test datasets	N_{valid}/N_{test}	5
Feature dimension	M	48
Reduction method	-	pilot
No. of training samples	N_{train}	10
No. of samples per data set	N	10000
User velocity	v_B	$1\frac{m}{s}$

7.3 Results

The PLA performance of the NB classifiers is depicted in Figure 5 for different AIs. While the MNB and CNB have a more or less constant accuracy independent of the AI, the accuracy of the GNB decreases for low AIs. Such as at an AI of 20%, 99% accuracy is reached, whereas for AIs≥40%, the accuracy is ≥99.99% for all classifiers.

Within Figure 6, the ROC curves of NB classifiers are presented for low (a), medium (b) and high (c) AIs respectively. The PLA performance of the GNB is decreased here similarly for a low AI of 10% reaching an AUC score of 98.14%. While the AUC performance increases for the GNB with increasing AI to 99.56% (medium AI) and 99.47% (high AI) respectively, for the MNB and CNB a decrease of AUC performance is observed. Both are reaching AUC scores of 99.75% at medium AI and 99.06% at high AI.

The PLA performance in terms of accuracy depending on the feature dimension M and training set size N_{train} is depicted in Figure 7. In case of low AIs (Figure 7a), the PLA performance of the GNB is constant at 99.6% independent of M, while the performance of GNB and MNB is decreased for very low feature dimensions (M=4) reaching a minimum of 95.32%. However, for M=20, M=24 and M=48 it reached accuracies ≥99.98%. At medium AIs (Figure 7b), the minimum PLA performance reached 99.2% for the CNB and NMB classifiers and 99.9% for the GNB classifiers at M=4. For M≥16, all classifiers reach accuracies of ≥99.98%. A similar result is observed for high AIs (Figure 7c), where the minimum PLA performance reaches 99.85% at M=4 and 100% for M≥8 for all classifiers. In case of a decrease of the training set size, the PLA performance decreases for low (Figure 7d) and high (Figure 7f) AIs at low values of N_{train}. At $N_{train} = 50$ for instance, the performance of the GNB classifier reaches 94% and 98% respectively. In case of the GNB and MNB classifiers, a PLA performance of ≥99.8% is reached for Ntrain≥100 in this case. In case of a medium AI (Figure 7e) however, all classifiers reach a minimum performance of 99.99% in all cases.

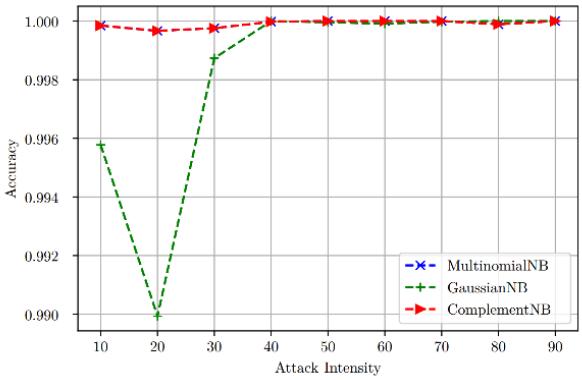


Figure 5: Accuracy scores for NB classifiers vs. Attack Intensity

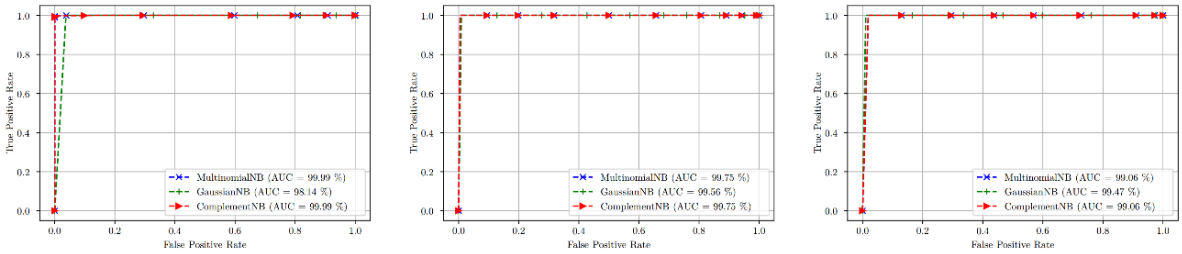
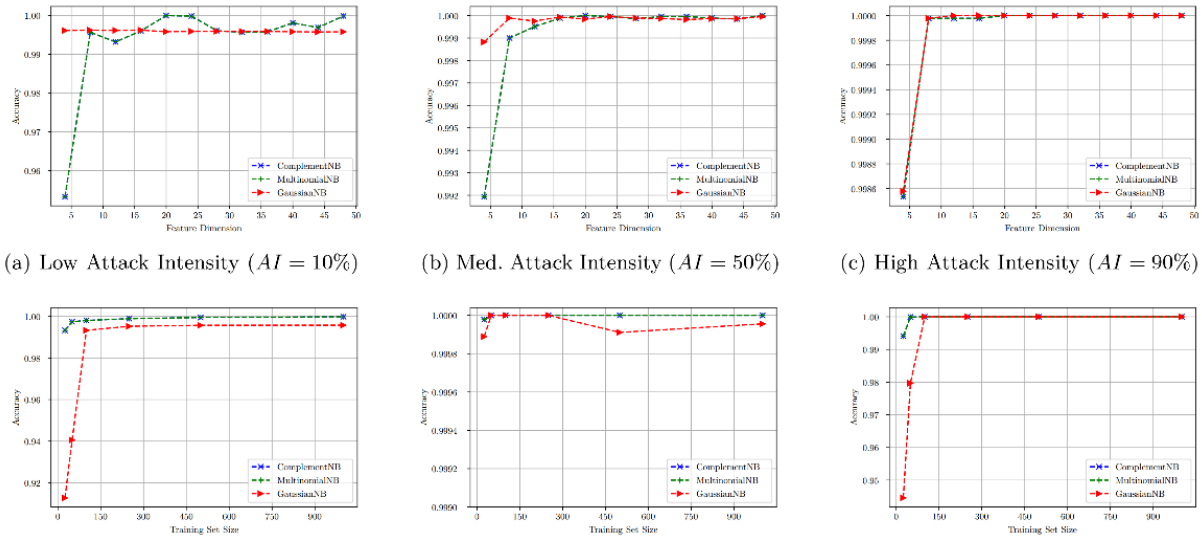
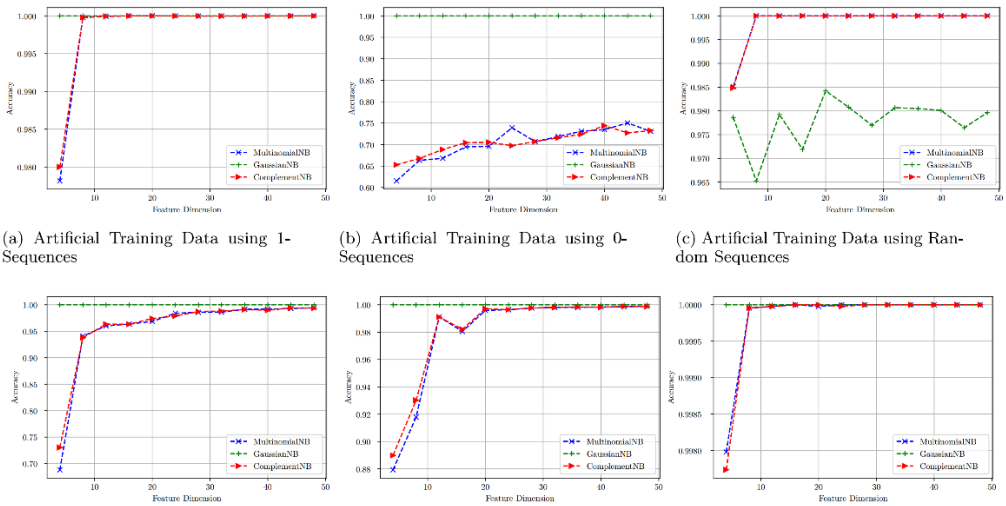


Figure 6: ROC curves for NB classifiers under different Attack Intensities



(a) Low Attack Intensity ($AI = 10\%$) (b) Med. Attack Intensity ($AI = 50\%$) (c) High Attack Intensity ($AI = 90\%$)

Figure 7: Accuracy of NB classifiers vs. feature dimension (a)-(c) and training set size (d)-(f) under different Attack Intensities (low, medium, high)



(a) Artificial Training Data using 1-Sequences

(b) Artificial Training Data using 0-Sequences

(c) Artificial Training Data using Random Sequences

Figure 8: Accuracy of NB classifiers vs. feature dimension under different Artificial Training Strategies, AI = 0% (a-c) and AI = 100% (d-f)

Figure 8 presents the results in case of missing training data from either Eve (AI=0%) or Bob (AI=100%). In case 1-sequences are used for artificial training data generation for Eves data (Figure 8a), the PLA performance is only decreased for low feature dimension (M=4) for the MNB and CNB classifiers at 97.2% and 98% respectively. In all other cases, the PLA performance is ≥99.98%. However, 0-sequences (Figure 8b) and random sequences (Figure 8c) provide suboptimal PLA performances for the MNB/CNB and GNB reaching a maximum of 73% and 98% at M=48 respectively. In case of missing training data from Bob, the PLA performance of the GNB is 100% independent of the used artificial training data generation method. However, in case of the MNB and CNB classifiers, the PLA performance decreases in case of a decrease in the feature dimension M. Minimum accuracies of 68.8-73% (1-sequences), 87.9-89% (0-sequences) and 99.77-99.8% (random sequences) are reached for instance at M=4. The maximum performance is reached at M=48 with accuracies of 99.36-99.45%, 99.87-99.9% and 100% respectively.

7.4 Discussion

The results presented in the previous subsection indicate a promising PLA performance in terms of the accuracy and AUC metrics. Further, the performance of the NB classifiers only decreases in rare scenarios (e.g. very low feature dimension or training set size). Especially the application of artificial training data strategies shows promising results in most of the cases. The choice of the respective method is however depending on the used classification scheme. In summary, the PLA performance derived within this work is similar or slightly above the performance reached in related work as presented in section 2.

8 Conclusion and Future Work

We have introduced the concept of PLA for industrial applications such as URLLC within this work. Further, we have proposed the usage of NB supervised classification for that purpose and presented methods in order to overcome the drawback of missing training data from respective users or attacker nodes in practical scenarios. The results derived from our experimental testbed show that under specific circumstances the application of NB supervised classification algorithms for PLA is a promising alternative to conventional security measures such as MACs. A considerable amount of radio resources can be saved within URLLC applications in that case. More advanced attack strategies such as e.g. burst attacks instead of attackers transmitting at a constant probability have to be considered within future work in order to further approve the robustness of PLA schemes.

Acknowledgement

A part of this work was funded by the Federal Ministry of Education and Research (BMBF) of the Federal Republic of Germany in the research project “Industrial Radio Lab Germany” (grant number 16KIS1011).

References

Baracca, P., N. Laurenti, und S. Tomasin. „Physical Layer Authentication over MIMO Fading Wiretap Channels." *IEEE Transactions on Wireless Communications*, July 2012.

Bockelmann, C., et al. „HiFlecs: Innovative Technologies for Low-Latency Wireless Closed- Loop Industrial Automation Systems." *22. VDE-ITG-Fachtagung Mobilkommunikation.* 2017.

Forssell, H., R. Thobaben, H. Al-Zubaidy, und J. Gross. „Physical Layer Authentication in Mission-Critical MTC Networks: A Security and Delay Performance Analysis." *IEEE Journal on Selected Areas in Communications*, 2019.

Gulati, N., R. Greenstadt, K. R. Dandekar, und J. M. Walsh. „GMM Based Semi-Supervised Learning for Channel-Based Authentication Scheme." *IEEE Vehicular Technology Conference (VTC Fall).* 2013.

IETF. „RFC 2104, HMAC: Keyed-Hashing for Message Authentication." 1997.

IETF. „RFC 4493, The AES-CMAC Algorithm." 2006.

McCallum, Andrew, und Kamal Nigam. „A comparison of event models for Naive Bayes text classification." *A comparison of event models for Naive Bayes text classification.* 1998.

Pei, C., N. Zhang, X. S. Shen, und J. W. Mark. „Channel-based physical layer authentication." *IEEE Global Communications Conference (GLOBECOM).* 2014.

Rennie, J. D. M., L. Shih, J. Teevan, und D. R. Karger. „Tackling the Poor Assumptions of Naive Bayes Text Classifiers." *Proceedings of the Twentieth International Conference on International Conference on Machine Learning.* Washington, DC, 2003.

Schmidl, T. M., und D. C. Cox. „Robust frequency and timing synchronization for OFDM." *IEEE transactions on communications*, 1997.

Senigagliesi, L., M. Baldi, und E. Gambi. „Comparison of Statistical and Machine Learning Techniques for Physical Layer Authentication." *IEEE Transactions on Information Forensics and Security*, 2020.

Tugnait, J. K., und H. Kim. „A channel-based hypothesis testing approach to enhance user authentication in wireless networks." *International Conference on COMmunication Systems and NETworks (COMSNETS 2010).* 2010.

Wang, N., T. Jiang, S. Lv, und L. Xiao. „Physical-Layer Authentication Based on Extreme Learning Machine." *IEEE Communications Letters*, 2017.

Weinand, A., R. Sattiraju, M. Karrenbauer, und H. D. Schotten. „Supervised Learning for Physical Layer Based Message Authentication in URLLC Scenarios." *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall).* 2019.

Xiao, L., L. Greenstein, N. Mandayam, und W. Trappe. „A Physical-Layer Technique to Enhance Authentication for Mobile Terminals." *2008 IEEE International Conference on Communications.* 2008.

Xiao, L., L. Greenstein, N. Mandayam, und W. Trappe. „Fingerprints in the Ether: Using the Physical Layer for Wireless Authentication." *2007 IEEE International Conference on Communications.* 2007.

Yilmaz, O. N. C., Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, und J. Sachs. „Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case." *2015 IEEE International Conference on Communication Workshop (ICCW).* London, 2015.

Zhang, H. „The Optimality of Naïve Bayes." *In FLAIRS2004 conference.* 2004.

How the Russian Influence Operation on Twitter Weaponized Military Narratives

Dana Weinberg[1], Jessica Dawson[2], April Edwards[3]
[1]Queens College-CUNY, Queens, NY, USA
[2]Army Cyber Institute, US Military Academy at West Point, West Point, NY, USA
[3]US Naval Academy, Annapolis, MD, USA

dana.weinberg@qc.cuny.edu
Jessica.dawson@westpoint.edu
aedwards@usna.edu

Abstract: Since 2016, Russia has engaged in a dedicated influence operation against the United States to exacerbate existing cleavages in American society and to undermine US national security. Although distinctly modern in its use of social media platforms, the current methods align with old Soviet doctrine using information warfare to gain a strategic edge over competitors. We examine Russia's use of military-related content and profiles in their influence operation on Twitter and, in particular, the strategic deployment of military narratives. Using data from Twitter's comprehensive data archive of state-backed information operations, we find that 12.14% of the 1,408,712 tweets in English from 2009 through February 2021 contain military-related content. In addition, of the 2,370 fake accounts on Twitter tied to the IRA and GRU, 148 were from accounts posing as military or military-adjacent, and these accounts posted 12.7% of the influence operation's tweets. Together, tweets containing military-related content or coming from fake military and military-adjacent profiles account for 22.6% of the tweets identified as part of the Russian influence operation on Twitter. The Russians used narratives related to veterans, particularly sacrifice narratives and post-Vietnam government betrayal of sacrifice narratives. Patriotic sacrifice narratives were used to gather and engage an audience and to legitimize and amplify the content and accounts. In contrast, betrayal of sacrifice narratives were weaponized to amplify and escalate divisive social issues by tying them to veterans' sacrifices. We conclude the Russians amplified existing military narratives in American culture and used fake American military profiles to wrap anti-government sentiment in patriotic trappings to exacerbate existing social divisions. Turning Americans against their government achieves Russian strategic goals of removing American influence abroad and allowing Russia to have greater impact on the levers of international power that serve Russian interests.

Keywords: influence operations, Russia, information warfare, military, narrative, veterans

1. Introduction

The American military community remains one of Russia's targets for concerted influence operations. In 2019, the Vietnam Veterans of America reported that significant efforts were being made by Russian-backed social media operations to target veterans and servicemembers from all service components — active duty, National Guard, and Reserves (Goldsmith, 2019b). These operations impersonated veterans and veterans' organizations, posing as trusted sources, to encourage engagement and ultimately to perpetrate espionage and financial exploitation on unsuspecting victims. Russian-backed Facebook pages and accounts used military-related narratives to convince people of their authenticity and to draw them to their fake causes. One page, "Veterans of Vietnam," used the logo and other brand attributes of the Vietnam Veterans of America and gained credibility by "sharing information about the Vietnam War and veterans' benefits" (Goldsmith, 2019a, p. 6). At the same time, the fake accounts promoted highly divisive content, such as anti-immigrant narratives such as "vets before illegals" (Goldsmith, 2019b).

Russia's influence operation was not limited to Facebook, nor was it limited to impersonating veteran service organizations. From 2018-2020, Twitter identified over 5,000 accounts associated with Russia and the Internet Research Agency (IRA) (Twitter Transparency Center, 2020) as part of Twitter's transparency operations. This paper seeks to examine the use of military and military-adjacent content and profiles by the Russian influence operation on Twitter. In particular, we examine the military-related narratives — narratives about the military community or promoted by fake military accounts — embedded in the content of the Russian influence operation tweets. This paper contributes to knowledge about cyber security by illuminating the role of military narratives in social cyber warfare.

2. Background

Since 2016, Russia has engaged in a dedicated influence operation against the United States to sow discord and undermine Americans' trust in their institutions (Department of State, 1987; 116th Congress, 2017). Russia's current influence operations involve "information warfare, psychological operations, and deception campaigns"

(Bagge, 2019, p. 23). Russia has long recognized the value in controlling and manipulating information (Soldatov and Borogan, 2017) and it draws from deep Soviet experience with information and influence operations (Andrew and Mitrokhin, 1999; Bagge, 2019; Gioe, Lovering and Pachesny, 2020). Russia-backed active measures in cyberspace have repackaged Soviet techniques for stoking already tense social divisions. Critically, they have rarely, if ever, created new social divisions, instead exacerbating existing gaps and seams in civil society (Pomerantsev, 2019; Jankowicz, 2020). Although distinctly modern in its use of social media platforms, current methods align with Soviet doctrine using information warfare to gain a strategic edge over competitors (Soldatov and Borogan, 2017; Bagge, 2019) and to create a potent fifth column of "useful idiots" (Gioe, 2018; Michael, 2019). Repackaging Soviet disinformation tactics for the Digital Age, Russian influence operations on social media employ *maskirovka* or deception as a strategic tool to "undercut [a population's] self-preserving mechanisms and behavior" (Bagge, 2019, p. 23). Russia seeks to gain "reflexive control" over large swaths of the American public as witting or unwitting accomplices who act in ways that benefit Russia's strategic goals.

In this context, the specific targeting of the Vietnam Veterans of America by Russian influence operations on Facebook appears to be highly strategic. The Vietnam War represents not only an historically divisive time in American history — the 1960s — but also a highly controversial conflict that Americans widely believe they were tricked into supporting (Gibson, 1994; Huebner, 2011; Belew, 2018). Narratives of government betrayal of American veterans gained traction in the 1980s in popular culture through anti-establishment, hyper violent novels, television, and movies (Gibson, 1994). They also found expression in political rhetoric, particularly on the American Right (Reagan, 1988; Rodman, 2019) in service of a political agenda to curb government influence.

The Russian influence operation on Facebook sought to exploit and exacerbate these conflicts by emphasizing veterans as victims of government incompetence or neglect, highlighting government betrayal of military heroes. It also used veteran personas and pages to spread divisive social messages. The influence operators claimed greater legitimacy for their fake sources by masquerading as veterans. If successful, these active measures on Facebook would serve Russia's interests by turning America against itself and undermining the ability of the United States to impact Russia's strategic goals abroad.

In this paper we examine the extent to which the Russia may have used similar influence operation tactics on Twitter. We consider the tweets and profiles that Twitter attributed to the Internet Research Agency (IRA) and the Glavnoye razvedyvatel'noye upravleniye (GRU) – Russia's military intelligence agency (Twitter Transparency Center, 2020). We hypothesize that the Russian influence operation on Twitter used military topics as a source of trust to advance and amplify divisive messages. In particular, we explore the Russians' use of military narratives, especially narratives of sacrifice and betrayal and how they tied these powerful cultural narratives to divisive political issues.

3. The Use of Narrative in Influence Operations

Researchers investigating influence operations have used a wide range of tools and methodologies to understand how content spreads online (Ferrara et al., 2016; Starbird, 2017; Beskow, 2019; Dupuis and Williams, 2020; Goldenberg and Finkelstein, 2020). These efforts have largely focused on keywords or trending topics. Some research has begun to evaluate how social media has enabled individual storytellers to influence outcomes in specific conflicts (Singer and Brooking, 2018), but little work has focused on the weaponization of narratives.

The term "narrative" means different things in different research traditions and disciplines. Drawing from the narrative research tradition in cultural sociology, we define narratives as meaning-laden stories —whether cultural narratives or broad cultural messages, personal narratives or "scripts of self" (Lamont et al., 2017), or narratives as a discursive form of rhetoric (Polletta et al., 2011). In this formulation, narratives are inherently social activities, helping shape the boundaries of identity and also communicating moral meaning.

The meaning of a narrative is socially derived and relates both to the identity of the storyteller and form of storytelling as well as to the audience's identity. Narrative meaning is allusive and shifts depending upon the sentiment expressed and elicited as the story is shared and the other narrative references with which it is relayed (Polletta *et al.*, 2011; Polletta and Callahan, 2017; Dawson and Weinberg, 2020). Narratives often include or reference other stories that the audience knows to signal the moral or meaning of the story, and they are more powerful when they do so (Polletta, 2009; Polletta *et al.*, 2011; Polletta and Callahan, 2017). Narratives also garner messaging power by striking an emotional chord with the audience. For example, nostalgia narratives often orient the audience to a glorified past or traumatic event as a means of reframing historical events and affirming social identity and group belonging (Maly, Dalmage and Michaels, 2013; Dawson and Weinberg, 2020).

Narratives thus provide a powerful basis for collective action (Ocejo, 2011) when they evoke strong emotions and feelings of identification.

In American civil religious discourse, military narratives, especially patriotic narratives of soldiers' sacrifice, are especially powerful (Marvin and Ingle, 1999; Gorski, 2017). Military narratives tap into nostalgia about national identity and patriotism (Skitka, 2005), as well as sacrifice and heroism. Core stories invoke the binary codes of civil society related to liberty and repression, sacred and profane, or good and evil (Alexander and Smith, 1993; Smith, 2010). As the heroes of these sacrifice narratives, soldiers consecrate the nation through their sacrifice of blood and service. Moreover, through honoring these patriotic American heroes and the cause of their sacrifice — the nation — the audience shares their sacred honor (Dawson and Weinberg, 2020) and affirms its own patriotism.

Following the Vietnam War, Americans sought new national narratives to make sense of the defeat and answer lingering questions about the rightness of the war. One prominent narrative, which Gibson calls the New War narrative, points to government betrayal of the good soldier by constraining military power as the proximate cause for the defeat in Vietnam and the seemingly wasted sacrifice of the American military (Gibson, 1994).

We hypothesize that, using the military as a source of sacred trust, the Russia deployed both traditional sacrifice narratives and government betrayal narratives to promote divisive messages.

4. Methods

4.1 Data and Sample

This paper uses data from Twitter's comprehensive data archive of state-backed information operations. We analyzed the 1,408,712 English tweets from 2,370 accounts in this archive, with dates ranging from 2009 through February 2021. These accounts were identified by Twitter as belonging to the IRA and GRU influence operations (Twitter Transparency Center, 2020).

To detect military content in this archive, we focused on words that would appeal to military and military-adjacent audiences; would signal membership of the user in the military community; or would highlight members of the military as heroes, villains, causes to celebrate, or victims. We created a list of common terms and phrases related to military themes, messages, and hashtags that were particularly prominent on social media. We then crowdsourced a list of military-related terms from members of the military community active on Twitter. After removing common abbreviations, such as "Col" for colonel or "Gen" for general, which frequently appear in other English words and resulted in a high false positive rate, the list contains 571 entries for identifying military content as well as six phrases, such as "New England Patriots," which appeared frequently and are not military-related. The list comprises obvious terminology, for example, "veterans," "military," "green beret" that would be recognized by a wide audience as pertaining to the military as well as terminology recognizable mainly to military insiders, for example, "gunny", "2lt", "snake eater," and names of military bases.

TextQ was used to pre-process and filter the data (Edwards *et al.*, 2021). The TextQ platform performs exploratory data analysis of social media material, emphasizing keyword search. The search is done via a two-step process, the first to eliminate potential false positives (i.e., removing tweets with the term "New England Patriots" so that those tweets would not inadvertently match on a search of "patriot"), and then a second pass to reduce the corpus to tweets that contain the desired terms. TextQ will also provide an inverted filter, which will produce a corpus containing all tweets which are not selected for comparison purposes, although that feature was not used in this analysis. When applied as a filter, our military keyword list resulted in a corpus of 318,843 tweets containing military-related keywords or phrases.

4.2 Analysis

We ranked our sample of military-labelled tweets by the number of retweets, a measure of either message resonance (McDonald, Bail and Tavory, 2017) or of message promotion and dissemination by Russian agents. We selected a sample of the 5,000 most frequently retweeted messages containing military content or tweeted from military profiles respectively (10,000 tweets total).

Focusing on text references to veterans (including derivations on the term as described below), a key indicator of military sacrifice or of government betrayal narratives, we conducted qualitative content analysis of the narratives in this subset of messages. When conducting the content analysis, we considered the target audience

for these narratives as well as the villains, threats, and causes they describe in order to determine the underlying message communicated when used separately or in combination with other narratives (Dawson and Weinberg, 2020). We then investigate how military narratives were deployed strategically to attract, hold, and direct audience attention as well as how they were weaponized against particular groups and issues (Weinberg and Dawson, 2021).

5. Results

Military-related terms occur in 12.14% of the 1,408,712 tweets in English from 2009 through February 2021. Of the 2,370 fake accounts on Twitter tied to the IRA and GRU, 148 were from accounts whose profiles contained military-related terms, and these accounts posted 12.7% of the influence operation's tweets. Together, tweets containing military and military-adjacent content or coming from fake military and military-adjacent profiles account for 22.6% of the tweets identified as part of the Russian influence operation on Twitter. Figure 1 shows the density curves for the military-related content and profile tweets compared to others over time. There were few military-related tweets from 2009 through April 2014, and then a steady rise through 2015 and 2016, at the height of the operation. In 2018, prior to the mid-term elections, there was a peak in Russia-directed tweet activity, with a proportionately larger spike in military-related tweets. There was a marked drop in tweet activity in November 2018 when Twitter published its first transparency report and deplatformed a number of suspect accounts.

Despite the depth of the list that we used for identifying military-related tweets and profiles, the most basic terms related to the military were the most prominent in the tweet texts in our sample (see Table 1). In contrast, the insider-based terms, such as those that referenced a specific unit or deployment did not appear, suggesting that while the Russians impersonated military personnel, they did not do so with any depth of military insider knowledge or insight. Variations on the word veteran were common—for example, "veterans" (3,081), "veteran" (1,348), "vets" (1,056), "deptvetaffairs" (258), "vet" (221), and "usarmyvet" (177). The term "veteran" and its variations appeared more than 6,141 times (using variations that appeared at least 100 times), making it the second most frequently used term next to "military" which appeared 8,903 times. In addition, there were multiple additional variations related to veterans, for example, "gulfwarvet123" (71), "codeofvets" (54), "brothervet" (31), "disabledvetspcsx" (30), "trump4vets" (28), "usmccmbtvet" (20), "navyvets4trump" (14), and "avetwithptsd" (14), along with numerous other permutations that appeared less frequently.

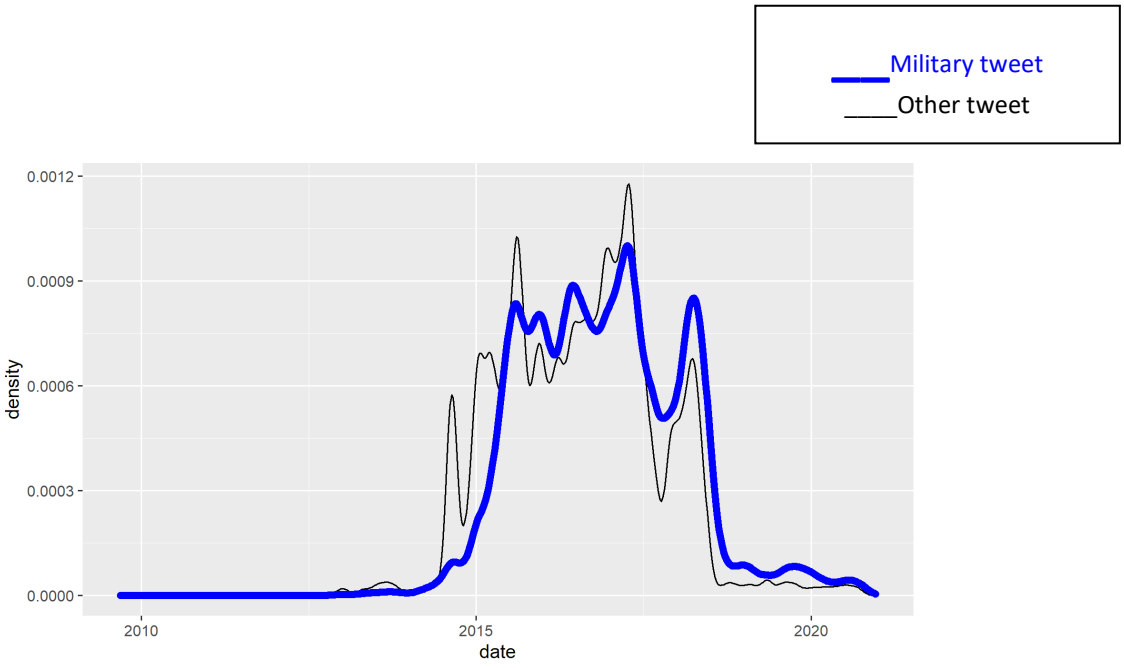


Figure 1: Density plot of military-related tweets in the Russian influence operation on Twitter, 2009-2020

Table 1: Top 5 Military Terms

Term	Number of Times Term Appears	Number of Documents Containing Term
military	8,933	8,606
army	4,812	4,727
veterans	3,081	2,990
troops	2,395	2,345
soldiers	1,996	1,962

The two most commonly identified narratives in our sample of retweeted posts were "veterans' sacrifice" and "betrayal of that sacrifice."

5.1 Sacrifice Narratives and Audience Gathering

The Russian tweets were consistent in their messaging about the military, emphasizing the heroic honor of soldiers and veterans, in line with the traditional military sacrifice narratives (Dawson and Weinberg, 2020), for example:

> Let's not forget that our peace and freedom are achieved by our veterans and protected by our soldiers. https://t.co/NBAQxk0zqA[1]
>
> The difference between the good guys and the bad guys is whether they use human shields or make themselves human shields. #VeteransFirst https://t.co/fhPfacLxRI
>
> We are eternally grateful to #Veterans. Their sacrifice will never be forgotten. Thank you! #BeingPatriotic https://t.co/vV1dMUJLyq

A significant portion of the military themed tweets served as audience building and highlighted feel-good stories about veterans that were used to elicit audience engagement in the form of likes and shares:

> Marine who lost leg in Afghanistan runs Boston Marathon with American flag. Give this PROUD PATRIOT some love!
>
> This guy was deployed in Afghanistan when he met a loving stray dog, and knew he had to find a way to get him back home. Touching. https://t.co/XByzFPVOTT

For the audience, liking and sharing these stories were presented as ways to show patriotism and support for American troops. Moreover, the audience was encouraged to give greater visibility to the account by retweeting its messages, thereby increasing organic reach and also building credibility and an audience for the fake accounts.

5.2 Government Betrayal Narratives and Weaponization

Many of the Russian military-content tweets informed audiences about specific veteran-related issues, including homelessness, healthcare, and PTSD and suicide. These tweets emphasized government's failure to honor the sacrifice that these issues represent.

> 50,000 US #veterans are homeless. Those heroes deserve better. #BeingPatriotic #VeteransDay #patriots https://t.co/YBEuY6A28s
>
> If we really want to pay tribute to our soldiers, we'd fix the military healthcare system https://t.co/QBzaqfnsiZ

Focusing on the shoddy treatment and suffering of the military heroes who sacrificed for the country, these tweets draw upon post-Vietnam narratives of government betrayal.

Going a step further, the tweets used narratives of government betrayal to spotlight divisive non-military issues, contrasting worthy veterans with those perceived as "less deserving". For example, tweets frequently

[1] We left the hyperlinks, though mostly defunct, to show where tweets referenced additional material.

contrasted narratives about the poor treatment of veterans with the services of dangerous or undeserving illegal aliens and refugees:

> Only in America rapefugees receive healthcare and 'housing' while Vets can't get anything!!! #AmericaFirst https://t.co/II2JY7NfXO

> Our vets deserve free education not illegal aliens who hate America and just want to get everything for free! https://t.co/8xtqyZVA2P

In this way, these accounts gained legitimacy with their target audience by showing support for the military while simultaneously associating the military with divisive, partisan political issues.

While most of the tweets we analyzed tied support for the military to right-leaning positions, a few linked military narratives to the interests of a left-leaning audience. For example, in the most-often-retweeted post with military content, the author took take aim at Trump in order to appeal to left leaning audiences:

> Don't ever tell me kneeling for the flag is disrespectful to our troops when Trump calls a sitting Senator Pocahontas, in front of Native American war heroes.

The tweet invokes outrage over Trump's derogatory remarks pertaining to the scandal surrounding Senator Elizabeth Warren's Native American heritage (Kaplan, 2019) in the context of a visit honoring the Navajo Code Talkers, who were widely credited with salvaging WWII efforts by using their native language as an unbreakable encryption (Bush, 2017). Moreover, with the reference to kneeling for the flag in protest of police brutality, the tweet emphasizes a narrative of sacrifice for a nation that has not lived up to its promises. This tweet was "liked" over 150,000 times and shared more than 50,000 times.

The Russia-based operation used betrayal narratives to turn sentiment against actors and institutions identified as unpatriotic. Beyond using veterans issues as a means of signaling supporting for or denigrating presidential candidates, the operation also used betrayal narratives to target the legitimacy of the American government and institutions:

> Govt has forgotten sacrifices of American heroes in #Vietnam. What could be worse for America? #BeingPatriotic https://t.co/YIPEilGDLi

> A LIVING NIGHTMARE! POLICE TAUNTED PARALYZED BLACK VETERAN AS HE LAY BEGGING FOR HELP AND DYING https://t.co/R7f5xfJGqz

> The government that doesn't respect its veterans should be overthrown! https://t.co/yGkl3CmtqA

The operation used government betrayal narratives to amplify negative sentiment toward various actors identified as unpatriotic and therefore illegitimate villains.

6. Conclusion

In this paper, we investigated the use of military content and profiles in the Russian influence operation on Twitter. We focused on the Russia's use of military narratives related to soldiers' sacrifice for the nation and post-Vietnam narratives of government betrayal (Gibson, 1994). We analyzed the strategic deployment of narrative to increase audience engagement and manipulate response.

We find that 22.6% of the approximately 1.47 million tweets from Russian-backed Twitter accounts either came from military profiles or included military-related content. The evidence suggests that the use of military narratives and profiles was a strategic maneuver, consistent with the goal of undermining trust in institutions, interfering with elections, and fomenting greater conflict around existing schisms in American society. The Russian influence operation used sacrifice narratives to gain and engage a patriotic audience and used government betrayal of sacrifice narratives to smear a variety of political actors, escalate the urgency of various social causes, and identify enemies, including the United States government.

6.1 Limitations

One limitation to this study is the use of exploratory NLP and the possibility both of false positives and false negatives. The use of abbreviations that commonly appear in other terms required the exclusion of abbreviated titles like “COL,” “GEN,” or “MAJ”; thus, the terms used to develop our filter for identifying military tweets was limited. As a result, we may be missing tweets that referenced a specific person with a military title, but no other military terms, but we believe the impact is negligible. It is not possible to identify all false positives or account for misspellings in a large corpus; thus, we may have missed tweets (false negatives) or inadvertently kept tweets (false positives). The almost limitless variety of approaches to conveying a concept in English make it impossible to eliminate these errors on large datasets. Given the emphasis on repetition and simplicity that seem to be hallmarks of the Russian influence operation, as well as the size of the corpus, we are confident that the false positives and false negatives do not substantially change the pattern of findings.

An additional limitation of this study involves assessing the impact of the Russian influence operation on individuals. Measuring influence requires a holistic view of the overall social media ecosystem as well as an understanding of how exposure to online content contributes to individuals’ attitudes and action. Developing measures to understand the effect of trending topics or repeated narrative exposure on social segments and individuals is an important area for further research.

6.2 Implications

Through the use of military sacrifice narratives and post-Vietnam narratives of government betrayal, Russia has leveraged existing sentiment of government betrayal of the troops to help drive deeper wedges between the republic and its citizens. We conclude that Russia used military narratives and profiles to wrap anti-government sentiment in patriotic trappings. Turning Americans against their government achieves Russia’s strategic goal to remove American influence abroad, allowing Russia to have greater impact on the levers of international power that serve Russian interests.

References to military narratives provide a means of narrative inflation, increasing the persuasiveness and mobilizing capabilities of narratives that utilize them. Creating the appearance of the military’s association with divisive partisan political issues further erodes Americans’ trust of the military as apolitical, while preserving support for the military as a marker of patriotism. As such, narratives that reference the military or that appear to be told by members of the military are a potent weapon in information warfare.

Acknowledgements

The authors gratefully acknowledge the New War Research Consortium and the Queens College-CUNY Sociology Department for their insights and feedback on earlier versions of this work. This research received funding under USMA grant W911NF-22-2-0042. This work does not reflect the views of the United States Department of Defense, United States Army, United States Military Academy, United States Naval Academy, or any government agency. Any mistakes are our own.

References

116th Congress (2017) Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election. Senate Report 116–XX. Washington, D.C.: United States Senate Intelligence Committee. Available at: https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf (Accessed: 23 January 2020).

Alexander, J.C. and Smith, P. (1993) ‘The discourse of American civil society: A new proposal for cultural studies’, Theory and Society, 22(2), pp. 151–207. Available at: https://doi.org/10.1007/BF00993497.

Andrew, C. and Mitrokhin, V. (1999) The Sword and the Shield: The Mitrokhin Archive and the Secret History of the KGB. 1st edition. New York: Basic Books.

Bagge, D. (2019) Unmasking Maskirovka: Russia’s Cyber Influence Operations. Defense Press.

Belew, K. (2018) Bring the War Home: The White Power Movement and Paramilitary America. Cambridge, Mass: Harvard University Press. Available at: http://www.hup.harvard.edu/catalog.php?isbn=9780674286078 (Accessed: 11 November 2018).

Beskow, D.M. (2019) ‘Social Cybersecurity An Emerging National Security Requirement’, Army University Press [Preprint], (Spring 2019). Available at: https://www.armyupress.army.mil/Portals/7/military-review/Archives/English/MA-2019/Beskow-Carley-Social-Cyber.pdf (Accessed: 14 October 2020).

Bush, el (2017) Families of Navajo Code Talkers decry Trump's use of 'Pocahontas', PBS NewsHour. Available at: https://www.pbs.org/newshour/politics/families-of-navajo-code-talkers-decry-trumps-use-of-pocahontas (Accessed: 8 November 2020).

Dawson, J. and Weinberg, D. (2020) 'These Honored Dead: Sacrifice Narratives in the NRA's American Rifleman Magazine', American Journal of Cultural Sociology [Preprint]. Available at: https://doi.org/10.1057/s41290-020-00114-x.

Department of State (1987) Soviet Influence Activities: A Report on Active Measures and Propaganda, 1986-87. Department of State Publication 9627. United States Department of State. Available at: https://www.globalsecurity.org/intell/library/reports/1987/soviet-influence-activities-1987.pdf (Accessed: 21 February 2020).

Dupuis, M. and Williams, A. (2020) 'Information Warfare: Memes and their Attack on the Most Valued Critical Infrastructure—the Democratic Institution Itself', 18(2), p. 11.

Edwards, A. et al. (2021) 'TextQ—A User Friendly Tool for Exploratory Text Analysis', Information, 12(12), p. 508. Available at: https://www.mdpi.com/2078-2489/12/12/508.

Ferrara, E. et al. (2016) 'The rise of social bots', Communications of the ACM, 59(7), pp. 96–104. Available at: https://doi.org/10.1145/2818717.

Gibson, J.W. (1994) Warrior Dreams: Violence and Manhood in Post-Vietnam America. Reprint edition. New York: Hill & Wang Pub.

Gioe, D.V. (2018) 'Cyber operations and useful fools: the approach of Russian hybrid intelligence', Intelligence and National Security, 33(7), pp. 954–973. Available at: https://doi.org/10.1080/02684527.2018.1479345.

Gioe, D.V., Lovering, R. and Pachesny, T. (2020) 'The Soviet Legacy of Russian Active Measures: New Vodka from Old Stills?', International Journal of Intelligence and CounterIntelligence, pp. 1–26. Available at: https://doi.org/10.1080/08850607.2020.1725364.

Goldenberg, A. and Finkelstein, J. (2020) Cyber Swarming, Memetic Warfare and Viral Insurgency: How Domestic Militants Organize on Memes to Incite Violent Insurrection and Terror Against Government and Law Enforcement. Princeton, N.J.: Network Contagion Research Institute, p. 11.

Goldsmith, K. (2019a) An Investigation into Foreign Entities Who are Targeting Servicemembers and Veterans Online. Vietnam Veterans of America.

Goldsmith, K. (2019b) 'Testimony of Vietnam Veterans of America: Hijacking our Heroes: Exploiting Veterans through Disinformation on Social Media'. Washington, D.C., 13 November. Available at: https://vva.org/wp-content/uploads/2019/09/VVA-Investigation.pdf.

Gorski, P.S. (2017) American Covenant: A History of Civil Religion from the Puritans to the Present. Princeton: Princeton University Press.

Huebner, A.J. (2011) The Warrior Image: Soldiers in American Culture from the Second World War to the Vietnam Era. Univ of North Carolina Press.

Jankowicz, N. (2020) How to Lose the Information War: Russia, Fake News, and the Future of Conflict. I.B. Tauris.

Kaplan, T. (2019) 'Elizabeth Warren Apologizes at Native American Forum: "I Have Listened and I Have Learned" (Published 2019)', The New York Times, 19 August. Available at: https://www.nytimes.com/2019/08/19/us/politics/elizabeth-warren-native-american.html (Accessed: 8 November 2020).

Lamont, M. et al. (2017) 'Bridging cultural sociology and cognitive psychology in three contemporary research programmes', Nature Human Behaviour, 1(12), pp. 866–872. Available at: https://doi.org/10.1038/s41562-017-0242-y.

Maly, M., Dalmage, H. and Michaels, N. (2013) 'The End of an Idyllic World: Nostalgia Narratives, Race, and the Construction of White Powerlessness', Critical Sociology, 39(5), pp. 757–779. Available at: https://doi.org/10.1177/0896920512448941.

Marvin, C. and Ingle, D.W. (1999) Blood Sacrifice and the Nation. Cambridge, U.K. ; New York: Cambridge University Press.

McDonald, T., Bail, C. and Tavory, I. (2017) 'A Theory of Resonance', Sociological Theory, 35(1), pp. 1–14.

Michael, G. (2019) 'Useful Idiots or Fellow Travelers? The Relationship between the American Far Right and Russia', Terrorism and Political Violence, 31(1), pp. 64–83. Available at: https://doi.org/10.1080/09546553.2018.1555996.

Ocejo, R.E. (2011) 'The Early Gentrifier: Weaving a Nostalgia Narrative on the Lower East Side', City & Community, 10(3), pp. 285–310. Available at: https://doi.org/10.1111/j.1540-6040.2011.01372.x.

Polletta, F. (2009) It Was Like a Fever: Storytelling in Protest and Politics. University of Chicago Press.

Polletta, F. et al. (2011) 'The Sociology of Storytelling', Annual Review of Sociology, 37(1), pp. 109–130. Available at: https://doi.org/10.1146/annurev-soc-081309-150106.

Polletta, F. and Callahan, J. (2017) 'Deep stories, nostalgia narratives, and fake news: Storytelling in the Trump era', American Journal of Cultural Sociology, 5(3), pp. 392–408. Available at: https://doi.org/10.1057/s41290-017-0037-7.

Pomerantsev, P. (2019) This is not propaganda: adventures in the war against reality.

Reagan, R. (1988) Remarks at the Veterans Day Ceremony at the Vietnam Veterans Memorial. Available at: https://www.americanrhetoric.com/speeches/ronaldreaganvietnammemorial.html (Accessed: 18 May 2020).

Rodman, R.C. (2019) 'Scripturalizing and the Second Amendment', in Y. Sherwood (ed.) The Bible and Feminism: Remapping the Field. Reprint edition. Oxford University Press.

Singer, P.W. and Brooking, E.T. (2018) LikeWar: The Weaponization of Social Media. Boston: Eamon Dolan/Houghton Mifflin Harcourt.

Skitka, L.J. (2005) 'Patriotism or Nationalism? Understanding Post-September 11, 2001, Flag-Display Behavior1', Journal of Applied Social Psychology, 35(10), pp. 1995–2011. Available at: https://doi.org/10.1111/j.1559-1816.2005.tb02206.x.

Smith, P. (2010) Why War?: The Cultural Logic of Iraq, the Gulf War, and Suez. 1 edition. University of Chicago Press.

Soldatov, A. and Borogan, I. (2017) The Red Web: The Kremlin's Wars on the Internet. Reprint edition. New York: PublicAffairs.

Starbird, K. (2017) 'Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter', in. Tenth International AAAI Conference on Web and Social Media (ICWSM), Montreal, Canada. Available at: http://faculty.washington.edu/kstarbi/Alt_Narratives_ICWSM17-CameraReady.pdf.

Twitter Transparency Center (2020) Information Operations - Twitter Transparency Center. Available at: https://transparency.twitter.com/en/reports/information-operations.html (Accessed: 1 September 2020).

Weinberg, D.B. and Dawson, J. (2021) 'From anti-vaxxer mom to militia men: Influence operations, narrative weaponization, and the fracturing of American identity', The Brookings Institution [Preprint], (October). Available at: https://www.brookings.edu/wp-content/uploads/2021/10/FP_20211108_influencer_operations_dawson_weinberg_v2.pdf.

Nuclear Weapons, Cyber Warfare, and Cyber Security: Ethical and Anticipated Ethical Issues

Richard L. Wilson[1] and Alexia Fitz[2]
[1] Towson University, Towson, MD
[2] Towson University, Towson, MD

wilson@towson.edu
afitz2@students.towson.edu

Abstract: In this paper, we discuss the interrelationship of nuclear weapons, cyber warfare, and cyber security. Some of the most significant cyber threats to nuclear stability are now due to the intersection of technologies related to nuclear weapons and cyber technology. Cyber warfare can now be used to engage in and influence international events through cyber attacks upon nuclear systems and weapons. In the current war between Russia and Ukraine there has been the threat of the use of nuclear weapons. Since cyber warfare has already been employed in the Russia/Ukraine conflict it is possible that cyber attacks could be employed to trigger a nuclear event. To prevent cyber warfare from leading to nuclear warfare there needs to be a focus on cyber security in order to protect nuclear systems and nuclear arsenals but also to mitigate cyber attacks that could lead to the use of nuclear weapons. One of the main risks to nuclear weapons systems is sabotage. It is easy to imagine cyber attackers placing incorrect information into systems and even taking control of nuclear weapons. Various parts of nuclear weapons systems are capable of being targeted. Command and control systems, alert systems, launch systems, and target-positioning systems could all become targets. Scenarios in which alert systems are hacked and show a nuclear attack by adversaries, may lead to an accidental nuclear conflict. It is also conceivable that hackers could manipulate the coordinates of (pre-programmed) targets of nuclear missiles, or to spoof GPS-like systems that some missiles use to calculate their positions their targets. At the present time there is no evidence that any state or non-state actor is able to successfully perform such manipulations but considering the exponential rate of developments in the cyber arena, in the near future, such attacks might be possible. In the worst-case scenarios, these possibilities could lead to the inadvertent use of nuclear weapons, and/or use against unintended targets. In less dramatic scenarios, the perceived vulnerabilities of the nuclear weapons systems may affect nuclear stability. This could lead to a decrease in the deterrent value of nuclear weapons. This could come about because potential adversaries may think they have options to manipulate these weapons when being used. It is difficult to forecast the effects of decreasing nuclear deterrence. This analysis will define a stakeholder framework for identifying the ethical and anticipated ethical issues with cyber warfare and nuclear warfare and relate these issues to the importance of cyber security. Ethics should be at the center of the discussion of the use of nuclear weapons, nuclear warfare and cyber warfare. Moral concerns should be at the center of the discussion of nuclear warfare. The need for this moral concern is due to the threat to vulnerable populations by nuclear systems and nuclear weapons, as well as the threat posed to democratic institutions by the use of nuclear weapons.

Keywords: Nuclear Warfare, Nuclear Weapons Systems, Cyber Warfare, Cyber Security, Anticipatory Ethics

1. Introduction

The issues that link nuclear weapons, cyber warfare, and cyber security are directly related to the notions of strategy and tactics. In the political and military arenas when nuclear weapons are involved, strategy and tactics play an important role. Military strategy, in nation state to nation state confrontations, is a major concern and focuses on how forces may be deployed against one's armed opponents. Strategy involves the planning, coordination, and general direction of military operations to meet a nations overall political and military objectives. Military strategy can be contrasted with military tactics. Tactics are involved with implementing strategy by short-term decisions on the movement of troops and employment of weapons on the field of battle. In discussions of traditional warfare, security can be thought of as developing, planning and coordinating military operations in order to protect one's nation. Once a conflict between nations commences, tactics implement a nation's strategy by the movement of troops and the deployment of weapons. Both nuclear weapons and technologies employed as weapons, and cyber technologies used for conducting cyber warfare, present challenges for traditional conceptions of strategy and tactics. An important reason for this is related to the differences of scale between weapons and conventional weapons and the complexity nuclear warfare and cyber warfare when they are used in conjunction. These challenges are evident at 3 levels. First, there is the issue of what is meant by nuclear strategy and what is meant by strategy in the context of cyber warfare? Second, what is meant by tactics in the context of nuclear strategy and warfare, and what is meant by tactics in the context of cyber warfare? Third, how is security to be understood in the context of political and military strategies and tactics as they relate to nuclear warfare and cyber warfare? These questions are central to the ethical and anticipated ethical issues with nuclear weapons, cyber warfare, and cyber security. Ethics is concerned with

principles that govern a person's behavior or the conducting of an activity. In the context of nuclear weapons, cyber warfare, and cyber security the stakeholders involved in ethical thinking about warfare and foreign policy, takes place within nations and between adversaries, and thoughts are directly related to intentions, actions, and outcomes related to each of these subjects which need to be identified in order to identify ethical issues with the use of nuclear weapons, cyber warfare, and cyber security.

2. Nuclear Warfare

The threat of nuclear warfare since the first of such weapon's, has been related to political strategy, and to the potential for the deployment of nuclear weapons. Nuclear weapons are weapons of mass destruction, which contrast with weapons employed in conventional warfare. Nuclear weapons are much more powerful and destructive in a much shorter time than conventional weapons and can have a long-lasting effects such as effects related to radiation. A major nuclear exchange would probably long-term effects, primarily from the radioactive fallout released by detonation, and could also potentially lead to secondary effects, such as nuclear winter, problems with famine, and the collapse of societies. Thermal thermonuclear war with Cold War-era stockpiles, or even with the current smaller stockpiles, may lead to various negative scenarios including the potentiality of extinction for the human race.(See: 7 Possible Toxic Environments Following a Nuclear War – The Medical Implications of Nuclear War 1985 – The National Academies Press. 1986.)

After the Soviet Union dissolved in 1991 and after the end of the Cold War, the threat of a major nuclear war between the U. S. and Russia have been considered to have lessened. More recently, concern over nuclear weapons has altered to the prevention of localized nuclear conflicts resulting from nuclear proliferation and the threat of terrorists getting access to nuclear devices. However, the threat of nuclear war is considered to have resurged due to the war in Ukraine, and threats by Russia to use nuclear weapons in the conflict. (Could the war in Ukraine go nuclear?". *The Economist*. 2022)

3. Effects of Nuclear Warfare

Nuclear warfare where nuclear weapons are employed produce a number of effects. The effects of a nuclear detonation on the environment are much more destructive and multidimensional than those caused by conventional weapons. The energy released from the detonation of a nuclear warhead detonated within the lower part of the earth's atmosphere can be approximately divided into four basic categories: (See: The Effects of Nuclear Weapons)

- The energy released in the blast itself: 50% of total energy
- There is the release of thermal energy: 30–50% of total energy
- Ionizing radiation is released: 5% of total energy
- Residual energy is also released: 5–10% of total energy with the mass of the explosion.

Depending on the design of a particular nuclear weapon, and the location in which it is detonated, the energy distributed to a target can vary. The environment of the explosion plays a large role in determining how much energy is distributed to the blast and how much to radiation. (See: The Effects of Nuclear Weapons).

3. Cyber Warfare

In nation state to nation state conflicts, cyberwarfare can be considered to be the use of cyberattacks against an enemy state, with the intention of creating damage that is similar to the damage created by conventional warfare. A cyberattack would be aimed at disrupting vital computer systems. In the context of this analysis the attacks would be aimed at disrupting nuclear systems and nuclear weapons including the detonation of a target nation's nuclear arsenal.

There is a debate among experts concerning the definition of cyberwarfare, and whether cyberwarfare exists. A number of nations in the international community including the U. S., the United Kingdom, Russia, China, Israel, Iran, and North Korea have extensive cyber capabilities for both offensive and defensive cyber operations. As more nation states develop the use of cyber operations and combine capabilities such as combining cyber and nuclear weapons, the likelihood of kinetic confrontation and violence developing as a result of the combination of cyber operations and nuclear operations is likely to increase.

What could be the first example of a kinetic military action which was used in response to a cyber-attack and that involved the loss of human life occurred on 5 May 2019. The attack attributed to the Israel Defense Forces targeted and destroyed a building associated with an ongoing cyber-attack (see: Newman, Lily Hay (6 May 2019). "What Israel's Strike on Hamas Hackers Means For Cyberwar"),

4. The potential effects of cyber warfare on nuclear warfare

There has been a great deal of discussion of nuclear weapons and Related Systems and how they are Increasingly Vulnerable to Cyberattack. **A** Nuclear Threat Initiative (**NTI) report (see:** https://www.nti.org/news/new-report-finds-nuclear-weapons-and-related-systems-increasingly-vulnerable-cyberattack/) has found that there are significant cyber threats to nuclear weapons and related systems—including nuclear planning systems, early warning systems, communication systems, as well as delivery systems. These cyber threats include an increase in the risk of unauthorized use of a nuclear weapon, an increase in the risk of nuclear weapons as a result of false warnings, and the undermining of confidence in the development of nuclear weapons as a deterrent. These threats are made possible by the speed, stealth, unpredictability, and challenges related to attribution with cyberattacks, all of which make it difficult to anticipate, deter, and defend against all cyber threats.

The NTI report also identifies the implications of these cyber threats to nuclear weapons, while making recommendations that governments need to employ to mitigate the threats that cyberattacks pose to nuclear weapons systems and nuclear weapons. These recommendations are stated in the report include:

- Developing options for increased decision time to account for threats to early warning systems
- Establishing norms that would restrict cyber weapons use against nuclear weapons systems
- Enhancing survivability and resilience of nuclear systems and command, control, and communications systems
- Securing and diversifying critical systems
- Prioritizing addressing cyber risks in modernization plans
- Maintaining teams of experts
- Enhancing security of nuclear weapons, and reviewing the vulnerabilities of nuclear weapons to combined physical and cyberattacks
- Initiating bilateral dialogue with Russia
- Increasing international cooperation to reduce the cyber threat.

5. Cyber Security

The previous recommendations made within the NTI report all point to the need for an increase in cyber security related to the risks of cyber warfare being used against nuclear systems and nuclear weapons. Computer security, cybersecurity (cyber security), or information technology security (IT security) are all concerned with the protection of computer systems and computer networks from attacks by adversarial and malicious actors. These attacks may result in the accessing protected information, theft of information or data, and damage to computer hardware and software. Cyberattacks could lead to the disruption and r misdirection of the services provided by computer networks. The threats posed by these types of cyberattacks to nuclear systems and nuclear weapons show the need for cyber security in these domains.

Cyber security has increased in importance as the result of the continual development and reliance on computer networks, the use of the internet and wireless network technologies. The continued development of technologies capable of connecting to the internet including Wi-Fi, Smart Devices, Smartphones and every device capable of being connected to the Internet of Things have lead to the interconnection of everything including nuclear systems and nuclear weapons. Cyber security faces significant difficulties for all aspects of the contemporary world, due to the reliance upon computer and information systems and the complexity of these systems. Cyber security is of significant important for systems that govern large-scale systems such as nuclear systems and nuclear weapons. The devices that connect to the emerging internet of things introduce multiple attack vector that could be employed to attack nuclear systems and nuclear weapons.

6. Nuclear Warfare, Cyber warfare and Cyber Security

There are a number of dangers that lead to the need for cyber security in the domains of nuclear warfare and cyber warfare. **Nuclear systems and weapons can be hacked.** Nuclear and information computer technology professionals are aware that all digital computer systems are at risk for cyberattacks which includes nuclear

weapons systems and nuclear weapons. This also means that any nation state that has warheads, delivery vehicles, and the technology used to control them, are capable of becoming prime targets and that they are vulnerable to cyberattacks and hacking. The defense systems relied upon to detect incoming attacks are also vulnerable. These same cyber vulnerabilities are in nuclear weapons systems and undermine the safety and security of nuclear weapons around the world.

The problem is that a successful cyberattack on a nuclear weapon system could result in nuclear detonation or even cause an all-out nuclear war. There are a number of ways that cyber threats could increase the chance of miscalculation and lead to nuclear use in the U.S.-Russia or Ukraine- Russia context (adapted from "The Cyber nuclear Threat" https://www.nti.org/analysis/articles/cyber/#:~:text)

1. Malicious actors are capable of interfere in U.S. or Russian or any nation's nuclear command and control systems.
2. Hackers could spoof the U.S. or Russian or any nations early warning system, leading either country to believe they're under attack.
3. A cyberattack on a communication system or nuclear systems and weapons could render the United States or Russia or any nation, unable to ensure that their nuclear weapons remain under proper control.

The vulnerability problem isn't confined to the use of nuclear weapons by the United States or Russia. Nine countries have nuclear weapons, and if these systems were compromised this creates a global nuclear risk.

7. Strategies for cyber security against cyber warfare on nuclear weapons

In another Nuclear Threat Initiative (NTI) report (Addressing Cyber-Nuclear Security Threats, see: https://www.nti.org/about/programs-projects/project/addressing-cyber-nuclear-security-threats/) a number of issues are presented related to cyber warfare and nuclear weapons. Important questions arise when we consider what kind of cyber threats could occur which are directed at nuclear systems and nuclear weapons. What would happen if a cyberattack shut down the security system at a highly sensitive nuclear materials storage facility, giving access to terrorists seeking highly enriched uranium required to make a bomb? Hackers could spoof a nuclear missile attack, which could trigger a miscalculated retaliatory strike that could kill millions.

The threat of cyberattacks can affect the nuclear risk environment in at least two ways: cyberattacks can be used to undermine the security of nuclear systems including the safety of nuclear materials and facility operations, and cyberattacks can compromise nuclear command and control systems for nuclear weapons. According to the NTI report nuclear security practices have in the past focused on preventing physical attacks—this means employing "guns, guards, and gates" to prevent 1) theft of materials to build a bomb, 2) sabotage of a nuclear facility, or 3) unauthorized access of nuclear command, control, and communications systems. Progress has been made in these areas, however the threat of a cyberattack against nuclear control systems and weapons is escalating. Nuclear cybersecurity practices need to be brought up to date with regard to the extend of current risks. The threat's to nuclear systems extends to the command, control, and communications (NC3) for nuclear weapons. Cyber threats are becoming increasingly sophisticated as information technology continues to develop and those responsible—from policymakers to military officials to facility operators to regulators have to continually work to keep up with these developments. There is a need to develop a set of guiding principles for cybersecurity at nuclear facilities. The current mindset needs to be replaced with the idea that there are ever-evolving threats to nuclear systems and nuclear weapons. An overarching framework that guides cybersecurity at nuclear facilities is needed.

8. BDI Model and Nuclear Cyber Security

The BDI model will serve as the foundation for understanding how actors involved in nuclear warfare, cyber warfare and cyber security are related to one another and eventually the distinctions from the BDI model can be employed to discuss how ethical issues emerge. What is BDI model? The Belief-Desire-Intention (**BDI**) **model** of human practical reasoning, developed by Michael Bratman (1987), is a model for assessing the rationality of human actions. The BDI model introduces future-directed intentions, which are composed to plans, as an important and irreducible concept. In the agent research community, the BDI model has been slightly adapted to specify the behavior of software agents in terms of beliefs, goals and plans. (Bratman, Intention, Plans, and Practical Reason). The model is diagrammed below.

Beliefs
→ intention → action → outcome
Desires

The BDI model plays an important role in predicting how agents will act and can also serve as the basis for predicting how agents will act, as well as for helping to predict how future directed reasoning will occur in practical reasoning which is directed towards achieving future goals. This model is used to predict the behavior of others while at the same recognizing we can apply the model to ourselves. The model can be employed to clarify how we develop our plans in order to attempt to achieve our goals. These beliefs and desires are the basis of intentions, actions and goals.

How can this model apply to nuclear weapons, cyber warfare, and cyber security? We will focus on political and military leaders in this preliminary analysis. There are nation states and political and military agents within them, who entertain the idea of nuclear war. These agents have beliefs and desires, they also have intention's and they have goals. There are agents who entertain the idea of cyber war. These agents who have beliefs and desires, they have intentions, and they have goals. There are agents who entertain the idea of cyber security. These agents who have beliefs and desires, they have intention's and they have goals. Adversaries in nuclear war and cyber war have beliefs and desires, they have intentions. and they have goals. As adversaries these are all capable of being at odds with one another. Security and Cyber security attempt to mitigate for example the intentions of adversaries to use nuclear warfare and cyber warfare against one another.

9. BDI Model and Ethics

How can the BDI model be related to ethics? For the purposes of this analysis we will remain focused on standard principles with philosophical ethics. Any number of principles could be employed for conducting an ethical analysis by relating them to an actor's intentions, actions and projected goals but here we will relate them to standard principles with philosophical ethics. Ethics in a basic definition relates to agents who perform actions. Dwight Furrow identifies and expands upon the focus of ethical analysis as involving a series of factors. As Furrow states, ethics is related to evaluating actions and actions are performed by those capable of being moral agents. Furrow continues by stating, "When we evaluate an action, we can focus on various dimensions of the action. We can evaluate the person who is acting, the intention or motive of the person acting, the nature of the act itself, or the consequences." (Furrow, Ethics: Key Concepts in Philosophy)

Two points are made within this passage. Ethical issues related to situations and cases are based upon the idea that when agents perform actions, these actions are capable of moral assessments. Moreover, actions are an extension of what a person does when they deliberate upon possibilities. In other words, the actions of agents are capable of being evaluated based upon both the intentions and actions of the person involved in the situation or case. If this is true and if we endorse the distinctions identified in the preceding passage and apply them to the assessment of moral responsibility in situations and cases, there are three levels of moral responsibility and evaluation. We can evaluate the actions of a person in the situation or case, we can assess the intentions of the persons involved in the situation or case, and we can assess the consequences of the actions intended by the person in the situation or case.

In order to understand the ethical issues involved with nuclear weapons, cyber warfare, and cyber security through situations and cases, we need to have an additional understanding of ethics and moral responsibility. Part of the difficulty related to ethical analysis requires the identification of what ethical principles to employ in our ethical analysis. Here we can only point to how this analysis would proceed.

Below is a preliminary diagram showing how ethical principles can be adjoined to Bratman's BDI model.

beliefs
goals
Character Traits Intentions → Actions → Outcomes
Virtue Ends
Ethics ↓ ↓
Desires

Deontology Consequentialism

Aristotle Kant and Ross Mill

Some Feminist
Ethicists

The distinctions on this diagram can be applied to the distinctions made above. Adversaries in nuclear war and cyber war have beliefs and desires, they have intentions, and they have goals. As adversaries they are capable of being at odds with one another. Security and Cyber security attempt, for example, to mitigate the intentions of adversaries to use nuclear warfare and cyber warfare against one another. Adversaries in nuclear war and cyber war have intentions (that violate or support rules and duties) and they have goals (that if achieved produce positive or negative consequences). A detailed ethical analysis apply ethical principles to each stakeholder perspective affected by nuclear weapons, cyber warfare, and cybers security.

10. Anticipatory Ethics

According to Phillip Brey (see: Anticipatory Ethics for Emerging Technologies, 2012) anticipatory ethics is concerned with examining ethical issues that arise from the emergence of technologies. This focus is particularly important due to the increasing rate at which technological developments are taking place. There is also an additional element that is important in this discussion, technologies not only continue to develop at a rapid rate, they are also converging with one another at a rapid rate. The focus of this analysis nuclear weapons, cyber warfare, and cyber security highlights difficulties related to the developments of technologies in all of these areas. While there are technical issues and ethical issues with developments in each of these areas, there are also technical and ethical issues due to the convergence of these 3 areas. New technologies are created when disparate technologies converge. Anticipatory ethics is an important method for attempting to identify ethical problems and potentially emerging ethical problems that might occur when technologies converge.

The convergence of nuclear warfare, cyber warfare and cyber security present a number of difficulties related to intentions, actions and outcomes that adversaries can employ against one another. However, a thorough anticipatory ethical analysis would seem to require much more than an analysis of the viewpoints of adversaries in a conflict. There are a wide variety of stakeholders who would be affected by any use of a nuclear weapon. These stakeholders would include political leaders, military leaders, deployed soldiers, designers of weapons, builders of weapons, individuals in the area where a nuclear weapon has been deployed, medical personnel, and global individuals. This is at best an extremely partial list of stakeholders who would be affected by the use of a nuclear weapon. The complexity of ethical issues becomes larger when we construct a list of stakeholders affected by cyber warfare and then move on to construct a stakeholder list of those involved with cyber security. Finally, a list of stakeholders needs to be constructed that identifies those affected when nuclear, cyber warfare and cyber security technologies converge. In all of these arena's we contend that actors have all of the items referred to in the BDI model. Actors have beliefs and desires, they have intentions, they perform actions, and they have goals that are aimed at producing outcomes. The ethical issues that emerge are related to how the stakeholders involved in nuclear weapons, cyber warfare, and cyber security are related to one another. The intention to employ nuclear weapons in an armed conflict can be examined from each stakeholder perspective. The decision can be examined from the perspectives of politicians, military leaders or civilians potentially affected by the use of nuclear weapons. Ethical issues and responsibilities can be identified from each stakeholder perspective.

11. Conclusions

Due to length constraints we now need to narrow our analysis to 2 stakeholder perspectives, political and military leaders. When considered from the stakeholder perspective of politicians/military leaders, nuclear warfare and the use of nuclear weapons presents a number of advantages and disadvantages for issues for political and military adversaries that would affect their intentions, actions and goals.

The advantages of nuclear weapons include (adapted from 17 Advantages and Disadvantages of Nuclear Weapons)

1. **From a border-based perspective they reinforce the idea of nationalism.**
2. **Nuclear weapons serve as a deterrent to a global conflict.**
3. **Nuclear weapons technology creates a bargaining chip for countries that need it.**
4. **Nuclear weapons reduce the threat to a country's military forces.**
5. **Governments can position nuclear weapons in a variety of launch locations.**

6. **Nuclear weapons research has contributed to the creation of new technologies in other sectors.**
7. **The reliability of nuclear weapons is one of its greatest attributes.**

When considered from the stakeholder perspective of politicians/military leaders nuclear warfare presents a number of issues for adversaries that are also disadvantages.

1. **There will always be moral and ethical debates about the use of nuclear weapons.**
2. **Nuclear weapon detonations are directly connected to cancer development.**
3. **There are direct costs attributed to a government's nuclear weapons program.**
4. **Nuclear weapons devastate the environment.**
5. **The use of nuclear weapons creates a significant threat of terrorism.**
6. **The development of nuclear weapons creates hazardous waste.**
7. **Degraded delivery systems can cause a nuclear weapon to fail.**
8. **We create nuclear weapons from non-renewable resources.**
9. **It requires a specific skill set to develop or maintain nuclear weapons.**
10. **Everyone is still dealing with the after-effects of nuclear weapons testing.**

When we examine the advantages and disadvantages of nuclear weapons, the idea of a threat deterrent does not seem to aim at genuine peace. The development of nuclear weapons is a race to create something bigger or more defensive in an weapons development program is not aimed at ending the threat of nuclear war. The use of significant conventional weapons without the threat of nuclear detonation and radiation exposure and the severe loss of civilian life, would perhaps be a better way of achieving the same goals. If we approach the issue of the dangers of nuclear weapons from the perspective of the convergence of nuclear weapons, cyber warfare and cyber security, we can see that Cyber warfare presents a number of possible attack vectors for attacking all adversaries with nuclear weapon systems, nuclear weapons and cyber warfare.

In the case of nuclear warfare we now assume that we can use the distinctions referenced by Furrow that focus on the actions of agents and we can state that adversaries are concerned in terms of intentions, actions, and outcomes with exploiting all the weaknesses of adversaries. What anticipatory ethics must focus on is mitigating are the intentions, actions and outcomes aimed at by adversaries when nuclear warfare and cyber warfare are thought about together. Here we will focus on Cyber Security as the basis for accomplishing this mitigation. **The spread of new cyber capabilities combined with the increasing digitization of nuclear systems increases the likelihood of dangerous cyber-nuclear attacks focused on nuclear systems and nuclear weapons. (See: The Cyber Nuclear Threat,** https://www.nti.org/analysis/articles/cyber/#:~:text=**)**

The United States as well as other countries with nuclear systems and nuclear weapons, are concerned with outdated technologies that need to be replaced by modern digital components as part of broad nuclear modernization programs. However according to the U.S. Government Accountability Office, at least in the United States, these technologies could create more problems than they solve: there are "mission critical cyber vulnerabilities" in nearly all the weapons systems under development by the Department of Defense. (See: Nuclear Weapons Cybersecurity: NNSA Should Fully Implement Foundational Cybersecurity Risk Management Practices, https://www.gao.gov/products/gao-22-104195)

Due to these difficulties according to the GAO Cybersecurity Risk Management Practices need to be put into effect. According to the GAO analysis these practices include:

1. Cybersecurity roles and responsibilities for risk management need to be identified.
2. Organizations need to establish maintain and review risk management strategies for nuclear weapons systems.
3. Policies and plans for the cybersecurity program need to be documented and maintained.
4. Organization wide cybersecurity risks need to be assessed and updated
5. Designate controls that are available for information systems or programs to inherit.
6. Strategies for continually monitoring risks need to be developed and maintained across organizations.

Finally due to insights provided by the discovery of Solar Winds, Cybersecurity Risk Management Practices also need to be implemented in numbers of arenas. These areas include: (See: Nuclear Weapons Cybersecurity: NNSA Should Fully Implement Foundational Cybersecurity Risk Management Practices, https://www.gao.gov/products/gao-22-104195).

The **operational technology environment which** includes sub-contractors who manufacture equipment and who build control systems with embedded software to monitor physical devices or processes needs to be made cyber secure.

The **nuclear weapons IT environment** including IT in or in contact with weapons. NNSA has implemented or taken action consistent with implementing most of the practices in this environment and is developing specific guidance for contractors needs to be made cyber secure.

NNSA's cybersecurity directive requires contractors to oversee their subcontractors' cybersecurity measures, but contractors' efforts to provide such oversight are mixed, and three of seven contractors do not believe it is a contractual responsibility. The recommendation is that contractor and cub contractor environments need to be made cyber secure.

In addition to identifying the risk factors listed above, which involve technical issues related to risks for nuclear systems, nuclear weapons and cyber warfare, a thorough ethical and anticipatory ethical analysis would also need to identify all of the stakeholders potentially affected by nuclear weapons, cyber warfare, and cyber security. The difficulty of this task is related to the intentions, actions and goals of all of the stakeholders involved. Ethical issues for each stakeholder are related to their intentions, actions and outcomes. Ethical problems emerge when stakeholder intentions, actions and outcomes clash with other stakeholder intentions, actions and outcomes on important issues. This ethical analysis will require a sustained analysis of situations and cases related to stakeholder involvements with nuclear weapons, cyber warfare, and cyber security. Here we merely outline this extensive project. Here we have only outlined some of the problems that will require a much more extensive analysis. This will carried out in future analysis and research.

References

17 Advantages and Disadvantages of Nuclear Weapons, Future of Working, https://futureofworking.com/6-advantages-and-disadvantages-of-nuclear-weapons/.

Ackerman, Pascal. Industrial Cybersecurity: Efficiently monitor the cybersecurity posture of your ICS environment, 2nd Edition 2nd ed., Packt Publishing; 2nd ed. edition (October 7, 2021).

"Addressing Cyber Nuclear Threats", Nuclear Threat Initiative 2022, https://www.nti.org/about/programs-projects/project/addressing-cyber-nuclear-security-threats/.

Bratman, Michael. Intention, Plans, and Practical Reason, Center for the Study of Language and Information; New edition (March 1, 1999).

Brey, Phillip. Anticipatory Ethics for Emerging Technologies, Nanoethics, 6 (1):1-13 (2012).

"computer security | Definition & Facts | Britannica". *www.britannica.com*. Retrieved 12 July 2022.

Could the war in Ukraine go nuclear?". *The Economist*. Sept 29. 2022. https://www.economist.com/international/2022/09/29/could-the-war-in-ukraine-go-nuclear.

Cunningham, Chase. Cyber Warfare – Truth, Tactics, and Strategies: Strategic concepts and truths to help you and your organization survive on the battleground of cyber warfare, Packt Publishing (February 25, 2020)

Ehrlich, P. R.; Harte, J.; Harwell, M. A.; Raven, P. H.; Sagan, C.; Woodwell, G. M.; Berry, J.; Ayensu, E. S.; Ehrlich, A. H.; Eisner, T.; Gould, S. J.; Grover, H. D.; Herrera, R.; May, R. M.; Mayr, E.; McKay, C. P.; Mooney, H. A.; Myers, N.; Pimentel, D. & Teal, J. M. (1983). "Long-term biological consequences of nuclear war". *Science*. **222** (4630): 1293–1300.

Fitz, Alexia and Richard L. Wilson. Just Warfare: Is a Nuclear Attack an appropriate Response to a Cyber Attack? Forthcoming in the proceedings for the 18th International Conference on Cyber Warfare and Security 9 - 10 March 2023, Towson, Baltimore County, Maryland, USA.

Furrow, Dwight. Ethics: Key Concepts in Philosophy, Continuum, New York, NY. 2005.

Huskaj, Gazmend and Richard L. Wilson. **HUSUSUSKAJ, GUZMANDOFFENSIVE CYBERSPACE OPERATIONS AND ZERO-DAYS: ANTICIPATORY ETHICS AND POLICY IMPLICATIONS FOR VULNERABILITY DISCLOSURE,** JOURNAL OF INFORMATION WARFARE, VOL 20, ISSUE 1, 2019.

Huskaj, Gazmend and Richard L. Wilson. Anticipatory Ethics for Vulnerability Disclosure, published in the proceedings of the 18th European Conference on Cyber Warfare and Security, Dayton, Ohio, 2019.

Lin, Herbert. Cyber Threats and Nuclear Weapons, Stanford University Press; 1st edition (October 19, 2021).

Lucas, George. Ethics and Cyber Warfare: The Quest for Responsible Security in the Age of Digital Warfare, Oxford University Press; 1st edition (December 13, 2016).

Martin, Brian (December 1982). "The global health effects of nuclear war". *Current Affairs Bulletin*. **59** (7).

Newman, Lily Hay (6 May 2019). "What Israel's Strike on Hamas Hackers Means For Cyberwar" Wired, https://www.wired.com/story/israel-hamas-cyberattack-air-strike-cyberwar/.

"Nuclear Explosions: Weapons, Improvised Nuclear Devices". U.S. Department of Health and Human Services. 16 February 2008.

"Nuclear Weapons Cybersecurity: NNSA Should Fully Implement Foundational Cybersecurity Risk Management Practices". 2022, https://www.gao.gov/products/gao-22-104195.

Schatz, Daniel; Bashroush, Rabih; Wall, Julie (2017). "Towards a More Representative Definition of Cyber Security". *Journal of Digital Forensics, Security and Law*. **12** (2).

Schnur and Richard L. Wilson. Cold War Echoes: The Russian Effort to Interfere in the 2016 Election, published in the proceedings of the 13th International Conference on Cyber Warfare and Security, Dayton, Ohio, 2017.

Stoutland, Page O. PhD and Samantha Pitts-Kiefer. Nuclear Weapons In the New Nuclear Age, 2018. https://www.nti.org/wp-content/uploads/2018/09/Cyber_report_finalsmall_Zg5TarX.pdf.

Stoutland, Page O., Addressing Cyber-Nuclear Security Threats, 2022. https://www.nti.org/about/programs-projects/project/addressing-cyber-nuclear-security-threats/.

"The Cyber-Nuclear Threat", Nuclear Threat Initiative 2022, https://www.nti.org/analysis/articles/cyber/#:~:text.

The Effects of Nuclear Weapons, https://www.britannica.com/technology/nuclear-weapon/The-effects-of-nuclear-weapons.

"The Energy from a Nuclear Weapon". *www.atomicarchive.com*.

What is BDI Model. IGI Global Publisher of Timely Knowledge. https://www.igi-global.com/dictionary/bdi-model/2286,/.

Wilson, R. L. (2021, April), *Anticipatory Ethics as a Method for Teaching Engineering Ethics* Paper presented at 2021 ASEE St. Lawrence Section Conference, Virtual. Now in the ASEE proceedings at: https://peer.asee.org/38292 .

Wilson, Richard L. State Sponsored Information Deception, Social Media, and the Rise of 'Soft' Warfare, presented at the 13th International Conference on Cyber Warfare and Security, Dayton, Ohio, 2017.

Wilson, Richard L. State Hacktivism, Cyber Warriors and Cyber Warfare, presented at the 13th International Conference on Cyber Warfare and Security, Dayton, Ohio, 2017.

Robots Security Assessment and Analysis Using Open-Source Tools

Benjamin Yankson, Tyler Loucks, Andrea Sampson and Chelsea Lojano
HackIoT Lab - College of Emergency Preparedness, Homeland Security and Cybersecurity
University at Albany, Albany, USA

Byankson@albany.edu

Abstract: The Internet of things (IoT) has revolutionized many aspects of the world, including industrial systems, automobiles, home automation, and surveillance, to name a few. IoT has offered a multitude of conveniences for our daily lives, such as being able to control our thermostats remotely, view our home surveillance cameras while away, or have a smart television that can surf the web. However, the widespread adoption of IoT devices combined with their vast device vulnerabilities results in significant security risks reinforcing the need for more robust default security controls and public awareness. As such, this paper aims to discover and document security vulnerabilities in the Asus Zenbo Junior IoT robot, along with providing a few best practices when securing smart home devices. This work presents an experiment using several security vulnerability assessment tools such as Nmap and OpenVAS scans to assess cybersecurity vulnerability currently present on Zenbo based on the 4P forensic investigative framework. The result of the experiment shows multiple open ports were discovered, along with miscellaneous information that an attacker could use to their advantage to attack the Zenbo robot. Based on the result, this work presents various security precautions that can help users protect against cyber-attack.

Keywords: Security, Robot, Assessment, attacks

1. Introduction

With the number of IoT devices forecasted to nearly triple from 9.7 billion in 2020 to more than 29 billion in 2030, security in these types of devices needs to be a more prominent topic of discussion because of the enormous consequences that a compromised device can lead to (Vailshery, 2022). For example, compromised IoT devices such as a smart robot like Zenbo, smart television, or smart fridge could open a pathway in a network for an attacker to infiltrate through. Vulnerable IoT devices are a problem that can affect both everyday people and businesses of all sizes. The scale of IoT attacks can range drastically from someone spying through a webcam to turning a device into a zombie as a part of their botnet to take down corporate websites. IoT security is an issue that affects millions of people around the world, and some people or businesses may even be completely unaware that their device is compromised. In today's age, downtime or data breaches of any amount can amass severe monetary losses for organizations. A 2021 IBM report (IBM, 2021) demonstrated that the United States continues to record the costliest data breach incidents worldwide. The report depicted that the average global cost of a data breach has reached over $4 million, far greater than the Ponemon Institute's Cost of a Data Breach Report for 2020(Barati & Yankson, 2022); which put the average cost of worldwide data breaches in 2020 amounted to $3.86 million. To put this into context, between 2005 and 2019, a report from the Privacy Rights Clearinghouse (PRC) shows about 9,015 data breaches, accounting for 11,690,762,146 breached records (Demeyer, 2011). Remediation and repair efforts also have to be compensated for; that does not account for any losses from factors such as missed sales, which vary greatly depending on the type of organization (e-commerce, video streaming service, and others).

To begin addressing IoT security, we believe two specific aspects will assist in lowering the number of successful attacks. First, more extensive vulnerability testing from the manufacturer is required before the product ships. Two, increasing user awareness about IoT device security and best practices to implement into their home networks. Improving these two aspects will strengthen IoT security and, in turn, lower the rate of successful attacks across all sectors, including home, medical, and industrial systems. The result of our work will contribute to current academic and industry discussions by showcasing the various vulnerabilities and security issues present in our smart home devices, such as the Asus Zenbo robot. Secondly, this work intends to contribute to the discussion surrounding more robust security controls in home-based IoT devices to minimize the potentially devastating consequences that can follow a successful attack.

1.1 Problem Statement

Since the number of IoT devices in the world is in the tens of billions and cyberattack rates continue to increase year after year, IoT devices need to implement much stronger security controls (Vailshery, 2022). For example, according to Kaspersky, an anti-virus and computer security service provider, IoT cyberattacks doubled yearly during the first half of 2021. From January to June this year, some 1.51 billion breaches of Internet of Things (IoT) devices took place, which is an increase from 639 million in 2020. There are over 1.5 billion IoT breaches, most of which use the telnet remote access protocol (Cyrus, 2021). The failure to secure IoT devices has and will continue to have massive consequences for everyday users and businesses. Damages can come in the form of

data being stolen, privacy being violated, loss of current or potential customers, and so on. With cybercrime expected to grow 15% per year and possibly reach $10.5 trillion annually by 2025, there needs to be a larger focus on securing IoT devices around the world in hopes of helping prevent users and businesses from suffering a successful attack (Morgan, 2022). In other words, since the rate of cyberattacks and cybercrime will significantly increase over the next few years, this reinforces the need for IoT manufacturers and users to secure their products better.

While there is no surefire method to mitigate cyberattacks completely, manufacturers and users can take a couple of steps to secure themselves better. On the manufacturer's side, more comprehensive vulnerability assessments should be conducted to help discover any security issues that should be remediated before the product ships. Another step the manufacturer should take to strengthen security would be defaulting the device to automatically download and install software/security updates so that burden does not fall on the user. On the user's side, there are also a few steps to take to secure their IoT devices. One step is to place all IoT devices onto a network separate from the home network, such as a guest network. The other step is to ensure that the device is on the latest software/security patch version, even if automatic updates are enabled. Ultimately, this paper serves to showcase some of the vulnerabilities that can be found in smart home IoT devices like Zenbo. The rest of the paper is organized as follows: Section 2 presents an overview of the background and related work. Section 3 details the methodology, 4P forensic framework, and vulnerability assessment case study using Nmap and OpenVas. Section 4 presents a discussion of the experimental result, and Section 5 concludes the paper and highlights future works.

2. Background and Related Work

The Zenbo Junior robot is advertised as a home companion that packs "loads of IoT functions," which "helps you to be more efficient" by integrating more artificial intelligence into your life (Zenbo Junior—Expansive Applications, n.d.). Zenbo Junior is a smart robot that can follow commands from the user to perform many actions, such as following a user around, spinning in place, or controlling other home IoT devices through Amazon Alexa integration. ASUS also details how Zenbo can be an excellent tool for the medical field as Zenbo can act as an interface for medical patients to enter health care information. ASUS also offers "Zenbo Lab," which allows Zenbo to run the python code created. These use cases convey how helpful and convenient a smart robot can be for someone. However, while convenient, security is often an issue that is not touched on nearly as much as it should be. If access or control to Zenbo falls into the hands of an attacker, there is a possibility that the attacker could have access to many other devices. This can include other smart home devices ranging from security cameras to televisions or even baby monitors. At the same time, extremely convenient and helpful to many, security needs to be a top priority for owners of smart robots such as Zenbo because of the massive consequences that can result from a compromised robot.

In 2016 Mirai botnet used around 100,000 malicious endpoints, essentially made up of IoT devices (Woolf, 2016). In this case, the compromised IoT devices were part of the botnet controlled by two malicious actors who used these devices to conduct distributed denial-of-service (DDoS) attacks (What is the Mirai Botnet? n.d.). As a result, the attackers were able to take down many high-profile websites, including Twitter, Reddit, and Netflix, to name a few (Woolf, 2016). While the total monetary damage of a DDoS attack entirely depends on the organization's size and downtime, many sources estimate that the monetary damage can range anywhere from tens of thousands to hundreds of thousands of dollars (Newman, 2021). On top of that, experiencing a successful DDoS attack will also cause reputational damage, which could deter current and potential future clients from using an organization's service or product. Regardless of the number of damages, no organization wants to experience a successful DDoS attack, be spied on by strangers through a webcam, or have their home IoT devices controlled by a malicious attacker.

Another excellent example of a compromised smart home IoT device is the Enabot Ebo Air smart robot. Much like ASUS' Zenbo, this robot is equipped with a camera, speaker, microphone, and wheels, allowing it to move around the house freely (Enabot, n.d.). Essentially, Enabot developed this smart robot so that the home could be monitored while away and so that it could remotely "connect with your loved ones" (Enabot, n.d.). Unfortunately, while all of that sounded great, it was discovered that these robots had the same hard-coded admin credentials (Hashim, 2022). This means that if an attacker can access the network that the robot is on, they will also be able to connect to Ebo via SSH because of this vulnerability (Hashim, 2022). Once inside, the attacker would have access to all of Ebo's functionality (view a live feed of the camera, listen through the microphone, and others .) and all information stored on the robot. Although the vulnerability was caught before shipping, this example conveys the potential severity of a successful IoT hack.

Nevertheless, while smart home IoT devices make for an excellent target for attackers, the medical and industrial sectors of the world also face instances of IoT devices being compromised and used maliciously. Imagine the robot used to bring pills daily to the hospital, spying on a user without the user's knowledge. It is a scary and unsettling thought to think about it. In the case of the Aethon TUG smart robot, these fears came to fruition. These robots are used in hospitals to deliver medication and supplies and perform miscellaneous tasks. However, some of the vulnerabilities found in these robots could allow attackers to disrupt the functionality of the robots, surveil patients and even obtain confidential information through the camera (Lakshmanan, 2022). Also, attackers could even have launched ransomware through these robots' network connectivity which could have affected the entire hospital (Lakshmanan, 2022). Like Enabot's Ebo, these vulnerabilities were caught before any attacker could exploit them (Lakshmanan, 2022).

On the other hand, industrial robots assist many different manufacturing processes. They are primarily used in big factories to help increase efficiency in product production. This includes many tasks, such as moving large objects/materials, welding, and assembling parts. The critical difference between the previously mentioned robots and industrial robots is that if a vulnerability is successfully exploited, it can lead to serious physical injury on top of monetary and reputational damage. Like the other IoT robots, the vulnerabilities stem from the robots being connected to the Internet. This Internet connection is the pathway that malicious actors use to launch their attacks. As Gaudin states, "Malicious hackers could get into a robot's controller system and adjust its actions, which could create a dangerous situation in the factory or enable the robots to build unsafe products on the production line" (2017). Overall, no matter what type of IoT robot (smart home, medical, or industrial), there are multiple security concerns that each present a different challenge for manufacturers and users to manage.

Using a weak or out-of-date encryption standard will raise the chances of an IoT device being successfully breached, which could compromise important information stored on the device. These devices can also allow the patient to interact virtually with a medical professional, and if the information is being transmitted in plaintext, an attacker could conduct a man-in-the-middle attack to intercept the network traffic being transmitted (Langkemper, n.d.). Software such as Wireshark makes it incredibly easy to intercept and log any network traffic that is being transmitted. With medical patient information being protected under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the security and use of strong encryption of these robots are necessary to maintain the confidentiality and integrity of any medical information being transmitted. One of the best steps to take in protecting IoT devices is ensuring the use of advanced encryption standards (AES). AES is generally agreed upon as "one of the best encryption protocols available, as it flawlessly combines speed and security, letting us enjoy our daily online activities without any disruption." When used correctly, successful attacks can be prevented (Rimkienė, 2020). Ensuring that the data transmitted from IoT devices is encrypted is crucial to the overall security of IoT devices.

As it relates to IoT, access control is "a set of permissions for a connected camera or any IoT device that specify which users are granted access and the operations they are permitted to perform" (Access control list (ACL) for Internet of things (IoT), 2020). Only certain users are granted access to the system, which is something that many IoT devices should have when it comes to security and privacy. This concept benefits devices with cameras or microphones, such as baby monitors, smart robots, or home surveillance systems. Access controls can take many forms, including mandatory access control, role-based access control, and discretionary access control. While these forms of access control all operate and assign permissions differently, they all ensure that only the people who need access receive access. Implementing access controls on smart home IoT devices can ensure that users who have access can control and perform activities with the devices. This can mean anything from updating the device remotely to viewing the security cameras while away from home. Implementing access control is a common practice in the business world to maintain the principle of least privileges and ensure a more robust security posture. Those same principles can be applied to the home network with your own IoT devices, making your smart home IoT devices and network much more secure.

3. Methodology

To begin the security vulnerability assessment of Zenbo, we adapt the 4P Investigation framework as proposed by Yankson et al. (2020). The framework consists of planning, preservation, processing, and presentation of results. The process begins "Planning" component. This involves establishing the scope and identifying the Asus Zenbo Junior as the medium involved. The processing component allows the assessor to prepare, set up, scan the medium and store the information. This component requires the Investigator to select the toolset necessary to collect and analyze available information. For this experiment, we will be selecting Nmap and OpenVas. The preservation component allows the assessor to document the type of information available and the state. At

this stage, the assessor can gather the necessary information to understand the inner working of the medium under investigation. Finally, the Presentation component involves collecting and analyzing the information found.

3.1 Tool Setup & Scenarios

The investigation has been performed on the Asus Zenbo Junior smart home IoT(Figure 1) to acquire data such as media recorded by the Zenbo, OS information, OS patch level information, and User ownership information using Nmap and OpenVAS free vulnerability software. The vulnerability testing Zenbo Junior involved three pieces of equipment: Zenbo Junior, a WAVLINK router, a Macbook Air 2016 for Nmap, and a Dell XPS 15 7590 for OpenVAS. In addition, we chose to use free vulnerability scanners to search for vulnerabilities and system information (patch level, operating system version, and others.) to allow the replicability of our work.

We powered on the router but did not connect the ethernet cable to the wall jack to avoid connecting to the main campus network. Next, we connected Zenbo Junior and the laptop to the WavLink WiFi router depicted in Figure 2. The IP address of Zenbo can be obtained by either opening the" Zenbo Lab" or through the device's settings. The WiFi network of this Zenbo uses a class C IP address, in this case, 192.168.10.128. Considering that the laptop Zenbo connected to the router locally, we began scanning Zenbo for vulnerability and open ports. Figures 3-5 show that Nmap was used multiple times with various scans that searched for the device's operating system, open TCP/UDP ports, and MAC address.

The results show that Zenbo is vulnerable to attacks and can link the Zenbo App to show ownership information. Based on preliminary data and current research, we made the following hypotheses:

- H1: Internal storage content of Zenbo can be accessed or manipulated, putting into question the confidentiality and integrity of information.
- H2: Zenbo can easily be subject to attack due to multiple open ports

To experiment to prove our set hypothesis based on preliminary findings, we set up a DELL laptop running Macbook Air 2016 for Nmap and a Dell XPS 15 7590 for OpenVAS (Figures 1-2).



Figure 1: Asus Zenbo Junior Robot



Figure 2: WavLink WiFi Adapter

3.2 Nmap Scanning TCP Port on Asus Zenbo Junior

To conduct the initial Scanning to determine an open port that can subject the device to an attack, the attacker's machine-running interfaces must be active to access the interfaces of Zenbo and acquire port information. As per Figure 3 below, we scan using the command *nmap -p 1-65535 -T4 -A -v*. This command allows us to scan all TCP ports on Zenbo Junior. We repeated the same command for UDP. The TCP and UPD port scan is depicted in Figure 3, Figure 4, and Figure 5. The results show that the Zenbo Junior model ran on the Android operating system version 6.0.1 with the security patch level from August 1, 2018. Checking for updates on the device does not allow us to update the operating system version or the security patch level.

The scan, as depicted in Figure 3, found multiple open TCP ports open, including ports # 8422, #19321, and #19323. The scan result also identified the current Mac Address of the Zenbo. The UPD scan result shows approximately 13 more open ports in filtered states. The open UDP ports include #1060, #1485, and others. (refer to Figure 4 for a list of the result of open ports). There were also a handful of open TCP/UDP ports on

Zenbo. Specifically, UDP port 4444 was left open on Zenbo, and both TCP/UDP port 4444 have vulnerabilities associated with them. First, port 4444 is the default listener port for Metasploit (SpeedGuide). Also, “some rootkit, backdoor, and Trojan horse software open and uses port 4444” (McKay, 2021). McKay describes how malware can use this port to eavesdrop on traffic and communications, exfiltrate data, and even download new malicious payloads (McKay, 2021). Ultimately, information about Zenbo, such as its MAC address and any open TCP/UDP ports, was obtained by using various Nmap scans. Subsequent to this, we run OpenVAS against Zenbo junior to identify any known vulnerabilities. The operating system and patch level are outdated, which inherently introduces security concerns.



Figure 3: Nmap scanning for all open TCP ports

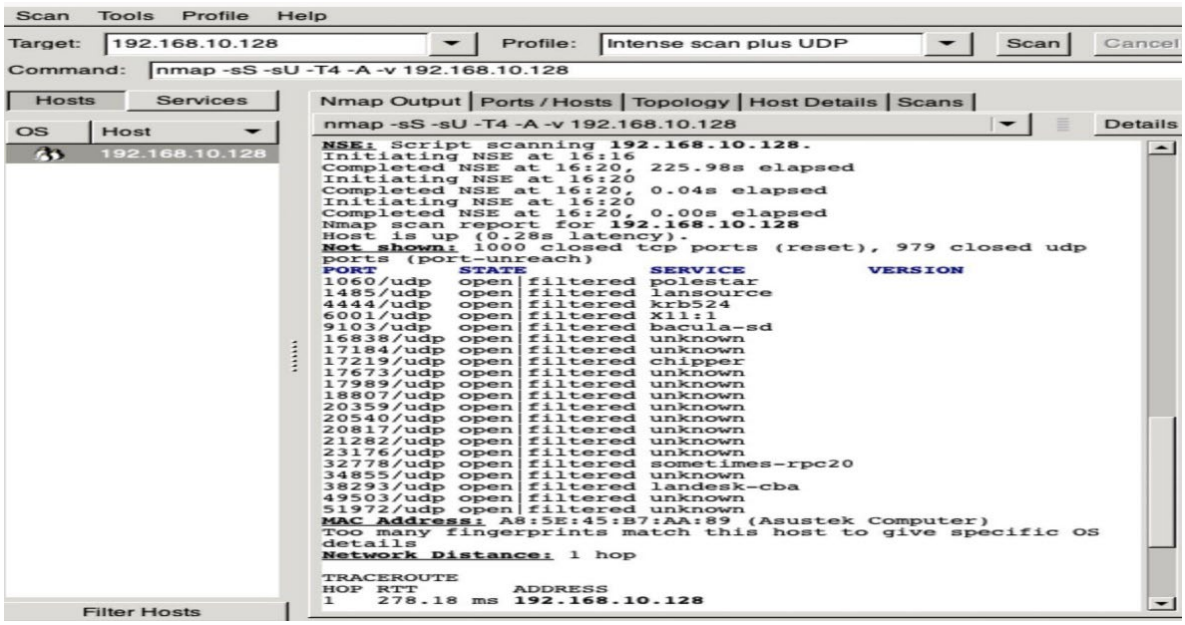


Figure @. Nmap scanning for all UDP ports

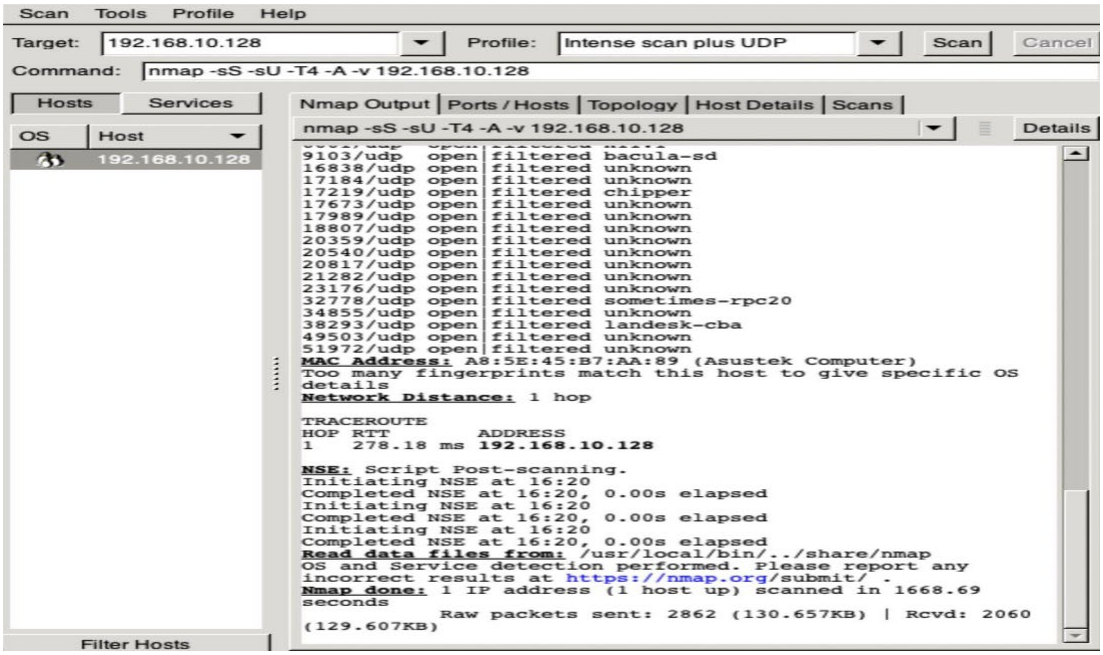


Figure 4: Nmap scanning for UDP ports as well as MAC Address

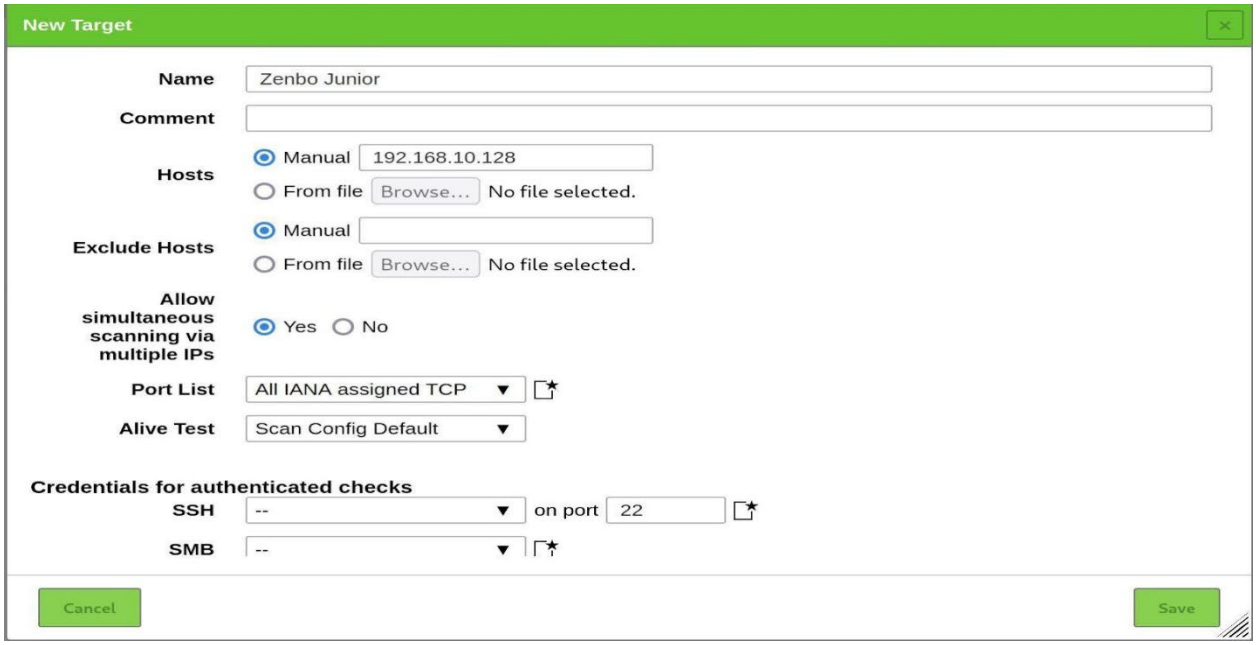


Figure 5: Task creation with OpenVAS

4. Discussions

The result of the experiment shows multiple open ports were discovered, along with miscellaneous information that an attacker could use to their advantage to attack the Zenbo robot. Since vulnerable home IoT devices can provide attackers a pathway into the home network, that also means that they could access any other network devices that are connected. This would include devices such as a Ring doorbell camera and Google Nest security cameras. This leaves the unsettling possibility of being spied on by a stranger. Such an act has been documented where an attacker could talk through one of the security cameras to a home user (Ring hacked, 2021). Unfortunately, many families worldwide have experienced this uncomfortable feeling of violating their privacy due to the weak security often present in their IoT devices. While the invasion of privacy is much harder to put a number on for damages compared to a successful cyber-attack, it can be argued that it is just as crucial for both organizations and users to consider the security and privacy implication when developing and purchasing smart home IoT devices, respectively.

Emerging IoT devices are inherently vulnerable to security and privacy issues which usually stem from exposed services/ports, weak encryption protocols, and lackluster access control. Regarding IoT security, the FBI states, "Your fridge and your laptop should not be on the same network. Keep your most private, sensitive data on a separate system from your other devices" (FBI, 2019). Separating the insecure IoT devices on your network ensures that even if an attacker compromises one of the IoT devices, they no longer have easy access to the rest of the network. Think of having a guest network for the home with IoT devices. In this case, the guest network would isolate these insecure devices from the rest of the network. Organizations implement different solutions, such as VLAN or DMZ, to achieve the same isolation concept with their public-facing web and email servers. While establishing and connecting the IoT devices to a separate network may seem arduous at first, it is well worth the time investment, as the alternative is accepting the possibility that the device becomes compromised by an attacker.

Another security precaution is changing the default credentials on the home devices. This step can help stop attackers from gaining access to the router, WiFi security cameras, and other wireless network devices. Attackers can easily exploit and gain access to a device left with default credentials, making this step an integral part of securing the IoT devices and overall network. Newly created passwords for the devices should contain more than eight uppercase and lowercase characters combined with symbols and numbers. Besides strong passwords, multi-factor authentication should also be considered, which provides much more security and something only the authorized user can access. Whether using SMS or an authenticator app to receive the one-time passcodes, this step greatly strengthens the security of devices. Lastly, since new security vulnerabilities are constantly being discovered, users must update their IoT devices regularly. Installing updates and patches can iron out any previously found security vulnerabilities that an attacker could have exploited. All in all, separating the IoT devices from the rest of the network, changing the default credentials on the devices, and maintaining the current patch level are some basic security steps that can help the user mitigate possible cyber-attacks.

5. Conclusions

There is no denying that smart home IoT devices device, such as Zenbo Junior robot, can offer users great convenience in their daily lives by assisting them with miscellaneous tasks such as remotely controlling their thermostats, wirelessly viewing their security cameras, or even being able to lock/unlock their front door while away from home. However, IoT devices such as these often suffer from inadequate security features, which can cause them to be a prime target for attackers attempting to infiltrate private networks. In this work, we have demonstrated that, based on the experimental results, the security vulnerabilities present in the Asus Zenbo Junior IoT robot. These vulnerabilities can make the Zenbo Junior robot an easy subject of attack due to multiple open ports. Also, the Internal storage content of Zenbo can be accessed or manipulated, putting into question the confidentiality and integrity of information. Based on the result, this work presents various security precautions that can help users protect against cyber-attack. Furthermore, the countermeasures provided can strengthen the security of their devices, such as changing the device's default credentials, keeping the devices up to date with the most recent patch level, and using a separate network, specifically the Asus Zenbo IoT device.

Based on the related work supports the finding that these vulnerabilities are not isolated to Asus Zenbo Junior. In addition, smart home IoT devices such as WiFi surveillance cameras or smart robots will usually suffer from similar security pitfalls, such as leaving unused ports open, shipping with an out-of-date security patch, or having lackluster authentication for remote access. These ever-present security issues can have massive consequences on the user and even cause damage to other organizations."

References

Access control list (ACL) for the Internet of things (IoT). AnyConnect. (n.d.). Retrieved November 4, 2022, from https://anyconnect.com/access-control/

Barati, M & Yankson, B (2022) "Predicting the Occurrence of a Data Breach." International Journal of Information Management Data Insights, Volume 2, Issue 2, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2022.100128.

Cyrus C. (2021), IoT Cyberattacks Escalate in 2021, IoT World Today. Retrieved December 20, 2022. https://www.iotworldtoday.com/2021/09/17/iot-cyberattacks-escalate-in-2021-according-to-kaspersky/

Enabot. (n.d.). Retrieved November 4, 2022, from https://na.enabot.com/

Demeyer, S. (2011) "Research methods in computer science." 2011 27th IEEE International Conference on Software Maintenance (ICSM), Williamsburg, VI, 2011, pp. 600-600. doi: 10.1109/ICSM.2011.6080841.

Gaudin, S. (2017, May 9). Industrial robots are a weak security link. CSO Online. https://www.csoonline.com/article/3195332/industrial-robots-are-security-weak-link.html

Hashim, A. (2022, July 28). Numerous Security Vulnerabilities Found In Enabot Ebo Air Smart Robot. Latest Hacking News | Cyber Security News, Hacking Tools, and Penetration Testing Courses.

https://latesthackingnews.com/2022/07/28/numerous-security- vulnerabilities-found-in-enabot-ebo-air-smart-robot/

Install Nessus Offline (Nessus). (n.d.). Retrieved November 14, 2022, from https://docs.tenable.com/nessus/Content/InstallNessusOffline.htm

IBM, *"IBM Report: Cost of a Data Breach Hits Record High During Pandemic."* (2022). Available: newsroom Online https://newsroom.ibm.com/2021-07-28-IBM-Report-Cost-of-a-Data-Breach-Hits-Record-High-During-Pandemic [Accessed: August 15, 2022] [Accessed: August 23, 2022]

Lakshmanan, R. (2022, April 15). JekyllBot:5 Flaws Let Attackers Take Control of Aethon TUG Hospital Robots. The Hacker News. Retrieved November 4, 2022, from https://thehackernews.com/2022/04/new-jekyllbot5-flaws-let-attackers-take.html

Langkemper, S. (n.d.). The most important security problems with IoT devices. Eurofins Cyber Security. Retrieved November 4, 2022, from https://www.eurofins- cybersecurity.com/news/security-problems-IoT-devices/

McKay, D. (2021, January 8). Why Are Some Network Ports Risky, And How Do You Secure Them? How-To Geek. Retrieved November 8, 2022, from https://www.howtogeek.com/devops/why-are-some-ports-risky-and-how-do-you-secure-them/

Morgan, S. (2020, November 13). Cybercrime To Cost The World $10.5 Trillion Annually By 2025. Cybercrime Magazine. https://cybersecurityventures.com/cybercrime-damages-6- trillion-by-2021/

Newman, S. (2021, October 17). The True Cost of DDoS Attacks. Infosecurity Magazine. https://www.infosecurity-magazine.com/opinions/the-true-cost-of-ddos-attacks/

SpeedGuide. (n.d.). Port 4444 (tcp/udp). SpeedGuide. Retrieved November 8, 2022, from https://www.speedguide.net/port.php?port=4444

Tech Tuesday: Internet of Things (IoT) — FBI. (2019, December 3). [Press Release]. Retrieved October 31, 2022, from https://www.fbi.gov/contact-us/field-offices/portland/news/press- releases/tech-tuesday-internet-of-things-iot

Rimkienė, R. (2020, December 11). What is AES Encryption, and How Does It Work?Cybernews. https://cybernews.com/resources/what-is-aes-encryption/

Ring hacked: How to protect your Ring smart device | NordVPN. (2021, December 23). https://nordvpn.com/blog/ring-doorbell-hack/

Vailshery, L. S. (2022, August 22). IoT-connected devices worldwide 2019-2030. Statista. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

What is the Mirai Botnet? (n.d.). Cloudflare. Retrieved November 3, 2022, from https://www.cloudflare.com/learning/ddos/glossary/mirai-botnet/

Woolf, N. (2016, October 26). DDoS attack that disrupted the Internet was the largest of its kind in history, experts say. The Guardian. https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet

Yankson, B. Iqbal, F. and P.C. K. Hung. (2020) "4P-based forensics investigation framework for smart connected toys". In Proceedings of the 15th International Conference on Availability, Reliability, and Security (ARES '20). Association for Computing Machinery, New York, NY, USA, Article 44, 1–9. https://doi.org/10.1145/3407023.3409213

Zenbo Junior—Expansive Applications, Flipped Learning. (n.d.). Zenbo Global. Retrieved October 27, 2022, from https://zenbo.asus.com/product/zenbojuniorii/overview/

Social Robots Privacy Enhancement Using Colored Petri Net (CPN) for Behavior Modeling: A Case Study of Asus Zenbo Robot

Benjamin Yankson[1], Farkhund Iqbal[2], and Fadya AlMaeeni[2]
[1]HackIoT Lab - College of Emergency Preparedness, Homeland Security and Cybersecurity University at Albany, Albany, USA
[2]College of Technological Innovation, Zayed University, Abu Dhabi, UAE

Byankson@albany.edu

Abstract: The interactions between a social robot and the user consist of continuous communication and behavior involving different data types that can be subject to cybersecurity attacks and prone to user privacy concerns. In this paper, we use Colored Petri Net (CPN) to develop two graphical models that illustrate different patterned behavior of a robot during such interaction. Using CPN, we model and analyze complicated robot system interactions considering synchronization and concurrency of events that can be subject to privacy violations. We focus on two specific scenarios involving user registration and medication reminder to provide an efficient illustration of the objects and events collaborated to carry out the intended tasks of the robot pertaining to privacy issues. The results show that CPN modeling simulation can capture and represent the robot's behavior, provide a better understanding of the task execution, and highlight users' privacy gaps requiring immediate controls.

Keywords: Privacy, Coloured Petri Nets, Modelling, Security, Robot

1. Introduction

For the last couple of years, one of the trending technologies has been social robots, which are widely used for numerous purposes and carry meaningful importance in social environments. Social robots are intelligent robots that interact socially with humans in their private environments, such as households or workplaces, and can mimic some human behavior. According to Taipale et al. (Taipale, 2020)], "Social robots are being designed to deal with human care, health, domestic tasks, entertainment, and various other immaterial and material tasks." These robots perform multiple activities based on their designs, making people feel comfortable and satisfied to interact with them. According to Tobe (Tobe, 2017), it is remarked that the robotics business is rising significantly, and it is predicted by the Boston Consulting Group (BCG) that the sales of robots "will reach $87 billion by 2025." (Tobe, 2017).

Similarly, Tractica predicted the sales of robots "will reach $237 billion by 2022." (Cole, 2020). Social robots are part of the Internet of Things (IoT) devices and comprise hardware and software modules with Internet connectivity, cloud services, and third-party applications for robot operations. Like any emerging computing technology, the amount of data residing in the robot constructions opens it to the same security and privacy concerns as other IoT devices. Moreover, the social robot's basic task is communicating with users visually and verbally; hereby, it continuously sends and receives data packets directly or over the network. The demand for Internet connectivity means that security holes would exist within the conceptual robot design because the fact is that anything connected to the Internet is prone to cyberattacks and subject to privacy violations (Poremba, 2016).

As such, the components of the robot system need to be understood to explore and identify all the possible security and privacy matters that a user may encounter during social robot interaction. Capturing the robot's behavior in a graphical representation to express the data flow between objects and describing the events would assist in identifying the vulnerable areas in the robot design and verifying the reliability of robot operations. To precisely capture the robot's behavior, formal models such as a Finite State Machine, automata, calculus, predicate logic, and Unified Modeling Language (UML) can be used. Accordingly, in this paper, we used Colored Petri Net (CPN) modeling mechanism to capture the behavior of ASUS' Zenbo robot during user interaction. CPN modeling is an extended version of the standard Petri Net (PN), which is a mathematical model applied to represent a Discrete Event System (DES), and it can be in two different structures: matrix representation or graphical representation (Da Mota, 2018). This model is widely used in computer science for software design, considering synchronization, sequence, and concurrency for complex system implementation (Reisig, 2013). CPN was introduced due to a significant limitation in the basic PN modeling: different data types cannot be referred to as a single state or object within the system. However, CPN allows users to define various data sets with different attributes. With CPN, the user can assign multiple attributes to one state depending on the requirements of the constructed scenario to assess privacy during user interaction simulation with the robot.

1.1 Security Implications

The logical examination of social robots' security problems demands understanding the robot's conceptual design and structure that integrates different components. Moreover, recognizing the services the social robot providential robot's services would help identify the features that the robot uses in communication and perform its functions, such as vision through gazing (gesture and motion), speech recognition, or screen display.

Many robotic systems are built on the ROS, which is used as middleware to provide several levels of software abstraction to hardware and robotics resources (i.e., sensor and actuators) in addition to reusing open-source project libraries (Koubaa, 2017). We surveyed different social robots to identify the platform used for robot development. It was observed that all of them are based on open-source platforms, including ROS and Linux. The vulnerabilities of Android and Linux-based operating systems are an ongoing research topic in the cyber security field because of their open-source feature and allowing for installing third-party applications on the system (Sankar, 2017).

Current research work and investigation have also identified a plethora of known vulnerabilities, including but not limited to robot interface authentication/authorization issues and bypasses and insecure transport of data (Zorz, 2017). Other avenues of vulnerabilities include a lack of firmware update mechanisms, improper and undocumented methods, hard-coded passwords, unencrypted storage, and easily disabled human safety protections (Zorz, 2017). Other than the identified vulnerabilities, some of these robots have exposed connectivity ports such as USB and Ethernet, with the possible remote access that a remote attacker can exploit to interfere with the robots' safety features. All such vulnerabilities can, in most cases, be exploited to violate users' privacy, hijack the robots, and in the worse cases, result in user safety in a case where the robot safety feature is compromised.

1.2 Privacy Implications

There are privacy issues associated with the use of social robots. Based on the services offered by social robots, users' sensitive data are collected and stored in their applications' databases or cloud storage. Since the user needs to be identified by the robot using different methods such as creating an account, face recognition, and voice recognition, personal details like name, email, and images are stored within the robot system. Moreover, the user interacts with different applications provided by the robot-like the reminder, calls, and information browsing. Therefore, users should consider their data protection critical while using robots for various activities. Such protection includes understanding all-important privacy issues such as robot manufacture used, robust cryptography methods to encrypt the local or cloud storage data, the ability of the user to identify and turn off location tracking and context data collection, the understanding of data ownership after transmission to the cloud, the availability of privacy policy and privacy compliance.

The contribution of this work is three-fold. First, We contribute to the literature by being one of the first to use Colored Petri Net (CPN) to model robot and user interaction to identify multiple data transition points that can be subject to a possible user privacy breach. In this work, CPN is used to develop two graphical models that illustrate different patterned behavior of a robot during such interaction. The presented model can be used to analyze complicated robot system interactions considering synchronization and concurrency of events that can be subject to privacy violations. The model capture and represent the robot's behavior, provide a better understanding of the task execution, and highlights users' privacy gaps requiring immediate controls. Second, this work highlights the privacy and security implications of humanoid robots. This work provides a better understanding of the task execution and the sequence of events. Third, CPN modeling is used to capture and represent the behavior of the Zenbo robot. We focus on two specific scenarios that demonstrate user registration on Zenbo and the capability of Zenbo to remind the user about his medication. The graphical representation of the situations would help simplify Zenbo's user interactions. The rest of the paper is organized as follows: Section 2 presents an overview of previous work. Section 3 details the methodology and implementation of a Zenbo scenario using CPN Tools. Section 4 concludes the paper and highlights future works.

2. Background and Related Work

This section provides an overview of previous work on social robots considering two sections: security and privacy and a Petri Net model.

2.1 Social Robot Security & Privacy

Chung et al. (Chung, 2017) discussed the complex operation of IoT devices, specifically the Amazon Alexa robot with Amazon Echo Wireless speaker, which they studied to identify the fundamental components that can be used as evidence in digital forensics crime cases. These components include cloud resources, user resources (mobile applications and web browsers), network resources, and hardware. Moreover, the researchers presented a systematic methodology called Cloud-based IoT Forensics Toolkit (CIFT), mainly built to obtain the inherent forensics objects from Alexa and analyze them using a Python script to extract forensics artifacts. The analysis includes different uncovered details, including the client files, Wi-Fi settings, Alexa-enabled skills, artifacts related to mobile apps, and a timetable for the results. Finally, Subramanian (2017) highlighted the significant implications of social home robots in several aspects, including security, privacy, and ethics. Additionally, he proposed some key points that can solve issues that several previous researchers raised.

Ray (Ray, 2016) provided an overview of the Internet of Robotic Things (IoRT) system. He described its design structure elements, consisting of different layers referred to as the Open System Interconnection (OSI) model. Besides, Ray examined the existing challenges in the system related to computation, optimization, security, and ethical issues. Do et al. (Do, 2018) developed a new robot called RiSH, Robot-integrated Smart Home, which aims to help older adults at home. The researchers presented the complete architectural design of the robot by describing all the hardware components used and the software components implemented. Besides, they built different services for the robot involving a variety of voice and motion sensors. The proposed RiSH system is believed to help other scientific researchers to conduct experiments on different robotic systems and highlight concerns related to the Information Technology field.

Portugal et al. (Portugal, 2017) inspected several home robots which use the ROS to analyze the most significant security concerns affecting the users, including authentication, authorization, and confidentiality issues. Furthermore, a case study is introduced to implement the possible security measures in each layer involved in a robotics ROS framework, including a wireless network, PHP Graphical User Interface (GUI), cloud server, ROS Multimaster, Operating System, disk data, and hardware level. The proposed protection system was called STOP Research and Development (R&D), mainly inherited from the actual STOP R&D Project.

Lee et al. (Lee, 2013) addressed data privacy associated with humanoid robots by conducting interviews with people to observe their understanding of data collection and protection within the robotic systems. The researchers highlighted the robotic features that pose privacy concerns, such as the various sensors built within the system, the motion capability, the verbal interaction between the user and the robot that reveals a lot of personal data, and sharing of social robots among different users. Fernandes et al. (Fernandes, 2016) also highlighted the risk of privacy issues of social robots since the camera is undoubtedly a part of the robot's systems. Thus, visual records can be obtained, revealing the users' unacceptable conditions. They discussed the privacy concerns specifically for older adults and disabled individuals because they are some of the primary users of social robots engaged within smart homes for their helpful capabilities. To enhance the privacy perspective in social robots and protect users' sensitive data, the authors presented a mechanism involving Convolutional Neural Networks (CNN): a concept used to examine images considering different features. In their work, CNN was used to recognize the private situations of the users where their privacy could be invaded. The model was built on the Advanced Sensing, Computation, and Control (ASCC) home service robot to navigate the social robots away from embarrassing situations based on its recognition using CNN. Moreover, they conducted a survey highlighting that people claimed their privacy and supported the implemented model.

2.2 Petri Net Model for Robots

Erden and Araz (Erden & Araz, 2016) presented a PN model developed to capture the walking behavior of a dog in real-time. It was built using the Artifex graphical modeling and simulation tool. The behavior in the model is represented by different functions, which can be used as transitions in the model. The dog's body parts (brain and four legs) involved in the behavior are referred to as places in the model. Andrews et al. (Andrew, 2016) represented Urban Search and Rescue Robots, where different disaster behavior models are described in the PN model. UML classes were used as a first step to describe each function and describe the functions' relations. Then, the PN model was built in reference to the created classes. Yang et al. (Yang, 2015) introduced a PN model integrated with Fuzzy logic, where the robot navigation functionality is used as a case to represent the PN model. It also explained the algorithm used to represent the PN model for the proposed scenario alongside the graphical representation. Two tools were used to build the Feature Pyramid Network (FPN) model: the HPSim Petri Net simulation tool and the YASPER tool. Chao and Thomaz (Chao & Thomaz , 2016) described implementing the PN

model for controlling real-time turn-taking decisions of the Simon Robot considering the time, called Timed PN. The paper guided the model through a building, mainly considering resources such as objects, spatial regions, the speaking floor, and the robot Degree of Freedom (DoF) represented by tokens and behavior (speech, gaze, gesture, and manipulation). Da Mota et al. (Da Mote, 2018) presented the PN model for a robot navigating behavior using RFID sensors by building four different paths. The model captures different actions: “stopping, moving forward, and turning right and left.” The authors' used matrix and graphical representations of the PN model. Another work that is instrumental in this area is the work of Stephenson (Stephenson, 2004). The author (Stephenson, 2004) proposes an optional method of risk analysis in Cybersecurity and quantification using statistical methods. Using proven formalisms, the authors used CPN for sophisticated modeling, simulation, and analysis of complex information security system behavior. The authors pointed out the advantage of CPNet's graphical representation, which helps construct, modify, and present complex models. Based authors' work, we build on modeling privacy spots in data transition for robot interaction.

3. Methodology - Petri Net Model Implementation

We analyze the two implemented CPN models and express the models in algorithm forms to clearly understand the sequence of events. The classical PN is composed of four fundamental components, where the fourth component is used only with CPN modeling, and they are (De La Mota, 2017):

1. *Places (P)*: denoted by circles, representing the system’s different states/objects/resources carrying out the events.
2. *Transitions (T)*: They are denoted by bars or rectangles, representing the events/interactions.
3. *Arcs (A)*: denoted by arrows, and they are used to represent the connections between places and transitions. It is impossible to draw an arc between the same type of nodes (i.e., between two places or between two transitions).
4. *Tokens* are denoted by dots and used to represent the data values inside each place based on the type of color set assigned to the place. These tokens are used as input data for the transition to be enabled and verify the action carried out. The tokens are added by marking the place.

3.1 Tool Setup & Scenarios

3.1.1 CPN Tools

Since We installed CPN Tools 3.4.0 on a Windows 7 VM running on VMWare Workstation to implement the CPN model (CPN, 2018). Although we use the ASUS Zenbo robot for this work, other researchers can use the methodologies to apply to other humanoid robots.

3.1.2 ASUS Zenbo Scenarios

The Zenbo robot is known to carry out multiple functions concurrent with different behavior interactions with the user. Our target in this work is to focus on a scenario to capture the robot’s interactions while executing a specific task involving a set of objects and events. We present the scenario captured through the CPN model, and for the scenario, we explain the developed model along with the involved events and objects. In addition, focusing on the security and general safety aspects, we highlight areas in the developed model that require end-user attention and awareness when using this social humanoid robot. Our presented scenario is about using the reminders App of the robot to add a medication reminder. The scenario was developed based on actual experiments on the Zenbo robot that involved studying some of the installed applications, and the way data is stored. Therefore, this approach was considered to create realistic CPN models based on factual data rather than assumed scenarios.

3.1.3 User Registration

Our first CPN model describes the process when users register themselves with Zenbo, as illustrated in Figure 1. Figure 1. shows that the process starts with P1, expressed by the STATUS data set to include the “OFF” token, which exemplifies that Zenbo is turned off. We have the declarations mentioned in Table 1. The “ZDATA” color set is composed of all the data types in the model, and the “STATUS” color set describes the different status types that an object can be at a given time. The initial User registration model comprises eleven places and ten transitions showing the execution of events sequentially. In addition, several tokens were defined based on the required input and output data for the states needed to fulfill the robot system interaction with the user. This model comprises eleven places and ten transitions showing the execution of events sequentially

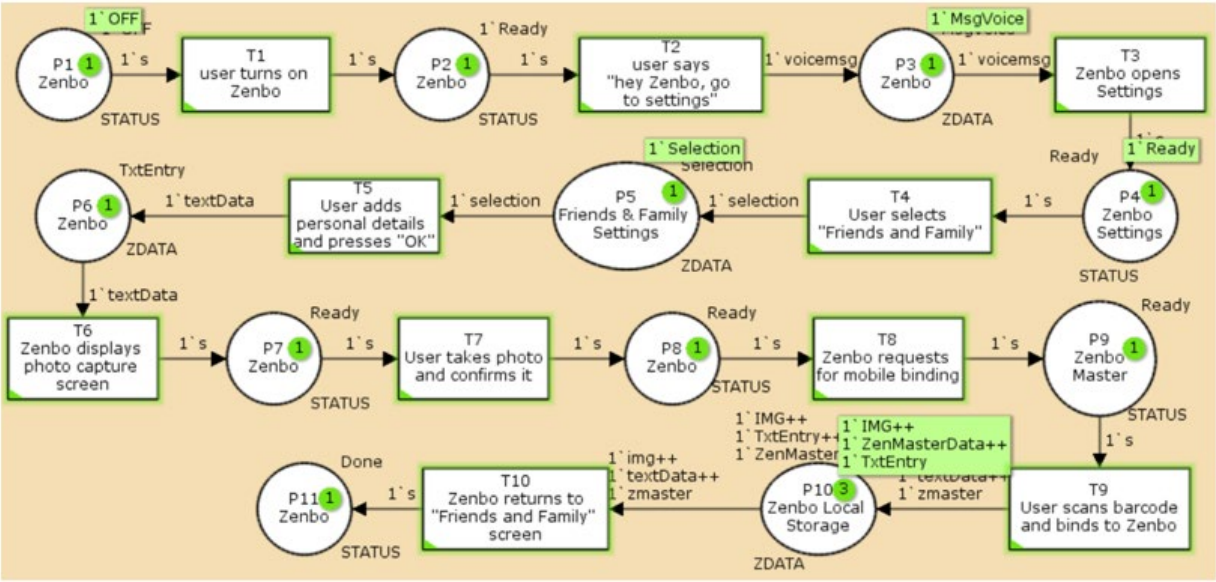


Figure 1. User Initial Registration

In this first model, we have the declarations mentioned in Table 1. The "ZDATA" color set is composed of all the data types in the model, and the "STATUS" color set describes the different status types that an object can be at a given time. We declared one variable referring to a STATUS color set, while five variables refer to the ZDATA color set.

Table 1. CPN Model 1 Declarations

Name	Type	Tokens/Values
ZDATA	colset	IMG, ZenMasterDate, TxtEntry, MsgVoice, MsgText, Selection
STATUS	colset	Done, ON, OFF, Ready
s	var	STATUS
voicemsg	var	ZDATA
selection	var	ZDATA
textData	var	ZDATA
img	var	ZDATA
zmaster	var	ZDATA

3.1.4 Reminder App

The reminder is a built-in app in Zenbo that provides valuable functionality for reminding users of different tasks through two options; a Calendar event or a To-Do-List item. This feature is vital for an older adult because it can remind the user to take the required medications at a specified time. For example, the CPN model illustrated in Figure 2 tells the scenario for adding a medication reminder through a To-Do-List item and then getting reminded by Zenbo at the specified time.

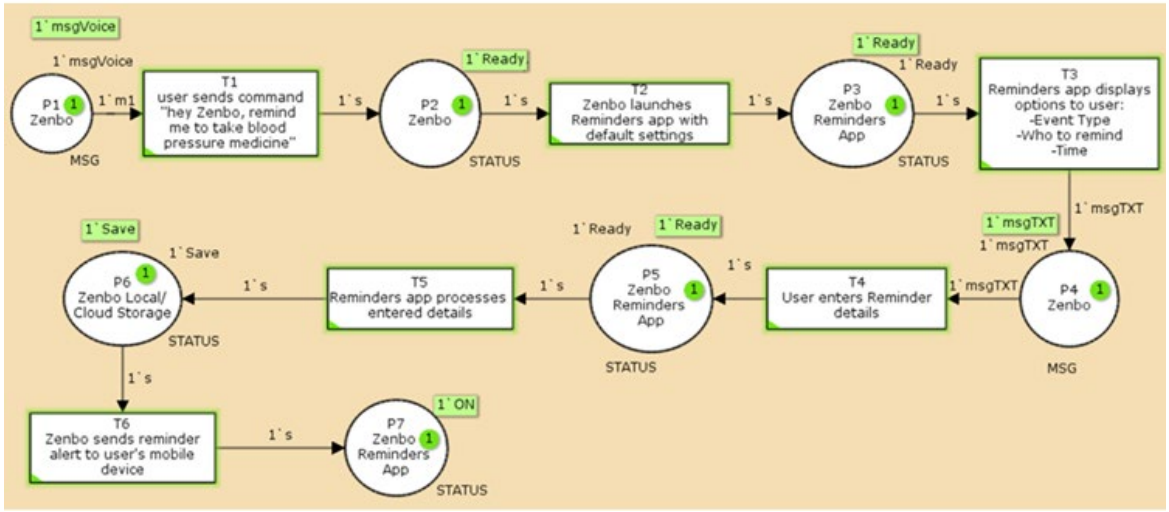


Figure 2. : CPN model of User Interaction with Zenbo Reminder app

3.1.5 Declaration

Table 2 shows all the color sets and variables declared for this model. This model declared three color sets with required tokens and two variables referencing some of these declared color sets.

Table 2: CPN Model 2 Declarations

Name	Type	Tokens/Values
ZDATA	colset	Msg, Voice
STATUS	colset	Save, ON, OFF, Ready
MSG	colset	MsgVoioce, MsgTXT
s	var	STATUS
m1	var	MSG

3.1.6 Model Description

Seven places represent the objects/resources involved in the process's execution of the Zenbo Reminders app used by the user. The significant entities of the structure include Zenbo, Reminders App, and the User. These entities are defined in several places characterized by one token or more having the same data types based on the case running to achieve the required task. However, six transitions capture the events executed sequentially between the places. They indicate the changes occurring between the states and the success of the data flow operation. For example, P1 represents Zenbo as it is ready to receive a voice message from the user. Zenbo gets alerted when the user initiates verbal communication with "Hey Zenbo" and posts these two words. The user continues by requesting the robot remind him to take blood pressure medication, represented by transition T1. There are different ways to create the required reminder on Zenbo; for example, the user may raise his request in one sentence to create the required reminder for a specific time and person, but in our model, we represent the process step-wise. Finally, P2 represents Zenbo again, but now it is ready to launch the Reminders App with the default settings, represented by transition T2, based on the initial voice commands sent by the user in an earlier transition.

Zenbo Reminders app opens and is ready for further action. This state is presented by P3, which triggers transition T3, which displays a list of other options to edit the reminder entry before saving it. Some fields the user needs to attend are the event type (calendar event or to-do-item), whom to remind among the members configured on Zenbo, and the time for the reminder to trigger. Based on our experiment, the to-do-item is a better option to select than calendar events in scenarios for medication reminders because such alerts get triggered on the Zenbo Masters app on the user's mobile device. While the user can configure the rest of the reminder entry options by voice commands, in our model, we present the configuration with a text/selection option since this is easier for the user based on our experiment. In transition T5, the Reminders app processes the reminder entry stored in P6, representing local or cloud storage. When it is time for the reminder to trigger based on the configured time, Zenbo launches the Reminders app that triggers an alert only on the mobile device linked to the robot via Zenbo's Masters' App. It is to be noted that though the originally published videos about Zenbo on the internet highlight that an alert for a to-do-item can be presented on Zenbo's screen, our

experiment did not find this possibility. The Reminders tutorial configured on Zenbo mentions that reminder alerts will be sent on a linked mobile device without mentioning the on-screen alert.

3.1.7 Observation

Table 3 lists the five observations that we have on this process.

Table 3: CPN Model 2 Observations

#	Description	Ref
1	Since it is difficult to set up voice recognition, as mentioned in the earlier model, this will mean that anybody nearby Zenbo can add a reminder to it. However, this may cause a medical issue for the elderly if two reminders were added for the type of medicine that is supposed to be taken once a day.	T1
2	The user cannot upload his medical record on Zenbo, so any entry for medicine reminders can be validated against the record. We recommend ASUS implement this enhancement on Zenbo since the robot can be used for elderly care as one of its stated purposes.	T1
3	It is challenging to have all the entries done through voice commands. So, the user may necessarily use selection instead of issuing voice commands. This makes it difficult for the elderly to set a reminder. ASUS could have made this process much easier by breaking the required entries into separate screens, with each screen displaying when the user can make direct voice commands to select or even choose to use manual entries.	T3
4	Reminder records are getting stored on the robot as well as syncing online. This could be proven by locating the extracted App's records that state the last syncing option and next syncing token. This is not a vulnerability, but it is worth highlighting since some users may be concerned about having their data on the cloud. Additionally, there is no encryption for the reminder data stored locally on the robot.	P6
5	The reminders App does not show the alert on the display screen of Zenbo. The ideal expected situation is that the alert can be displayed with a voice message to take his medicine. However, that does not happen. Instead, the user can get a reminder on the Zenbo Master mobile app, which may make it inconvenient for the elderly because not all of them are familiar with mobile devices, and they may not be as close to mobile devices as often as younger people.	T6

4. Analysis, Result, and Conclusions

The implemented models of Zenbo behavior may apply to any other social robot since any robot would have similar processes to execute its tasks. Using CPN modeling to capture social robot interactions with the user has several advantages from different perspectives. The user can design the complex system of the robot hierarchically, considering the synchronization of the workflow among the objects in detail. The graphical representation is highly comprehensive in that the user can easily assign captions to each model component, making the diagram readable. This feature allows the user to identify the needed resources for each task to be carried out and realize any system failures encountered while executing the intended task. In other words, the user can refer to the CPN model to follow the sequence of events and ensure that the robot is acting normally as expected based on the outcome of the robot. People need to consider the successful operation of the robots if they decide to depend on them for serious tasks.

Furthermore, CPN models may spot insight into the robot system's data privacy and security concerns. For example, the CPN model visualizes the data flow, representing the incoming and outgoing traffic through the system. This raises attention to exploring the expected security vulnerabilities that can exploit the operation of the robot processes. Likewise, the data collection and storage, which usually contain user data, must be ensured with high protection.

Social robots pose security and privacy concerns similar to any other computing smart device. This paper captured user registration, initial interaction with Zenbo, and Zenbo's reminders app for medication notification. The behaviors were represented using the CPN modeling tool, considering the tool's capability to deal with complex systems. Furthermore, the models were constructed to show the workflow of the intended behavior. As a result, CPN representations clearly understood the robot's interaction with the user and its internal systems, showing the process status with the flow of different data types.

Furthermore, the models allowed us to infer some hints about the robotic system's possible security and privacy issues. Finally, different recommendations were provided to overcome the robot forensic legal challenge and improve the security of robot systems. This work has provided insight into social robots' security and privacy implications.

References

Andrews, A., Abdelgawad, M., and Gario, A., (2016) "World Model for Testing Urban Search and Rescue (USAR) Robots using Petri Nets," Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development

Chung, H., Park, J., and Lee, S., (2017) "Digital forensic approaches for Amazon Alexa ecosystem," Digital Investigation, vol. 22, 2017, pp. S15-S25.

Chao C., and Thomaz, A., (2016) "Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration," The International Journal of Robotics Research, vol. 35, no. 11, pp. 1330.

CPN, T, (2018) "CPN Tools – A tool for editing, simulating, and analyzing Colored Petri nets",(2018) Cpntools.org,. [Online]. Available: http://cpntools.org. [Accessed: 10- Nov- 2018].[24]

Cole, E. (2019) "What is New in Robotics." Robotiq, [Online]. Available: https://blog.robotiq.com/whats-new-in-robotics-this-week-jul-14 [Accessed: 3-June-2022]

Da Mota, F., Rocha, M., Rodrigues, J., De Albuquerque, V., and De Alexandria, A., (2018) "Localization and Navigation for Autonomous Mobile Robots Using Petri Nets in Indoor Environments," IEEE Access, vol. 6, pp. 31665-31676.

De La Mota, I., Guasch, A., Mujica Mota M., and Angel Piera, M., (2017) "Robust Modelling and Simulation". Cham: Springer, 2017, pp. 49-52.

Do, H., Pham, M., Sheng, W., Yang, D., and Liu, M., (2018) "RiSH: A robot-integrated smart home for elderly care," Robotics and Autonomous Systems, vol. 101, pp. 74-92.

Erden Z., and Araz, M., (2016) "Petri Net Modeling and Simulation of Walking Behaviour for Design of a Bioinspired Robot Dog," Proceedings of the 6th International Conference on Simulation and Modeling Methodologies, Technologies and Applications.

Fernandes, F., Guanci Y., Do, H., and Sheng, W., (2016) "Detection of privacy-sensitive situations for social robots in smart homes," 2016 IEEE International Conference on Automation Science and Engineering (CASE).

M. Lee, K. Tang, J. Forlizzi and S. Kiesler, "Understanding users' perception of privacy in human-robot interaction," Proceedings of the 6th international conference on Human-robot interaction - HRI '11, 2011.

Portugal, D., Pereira, S., and Couceiro, M., (2017) "The role of security in human-robot shared environments: A case study in ROS-based surveillance robots," 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Petrosyan G., and Ter-Vardanyan, L., (2016) "Modelling of identification and secret-key generation system with Colored Petri Net," 2016 International Conference on Control, Decision and Information Technologies (CoDIT), 2016.

Poremba S., (2016) "The Internet Of Things Has A Growing Number Of Cyber Security Problems,"," Retrieved from http://www.forbes.com/sites/sungardas/2015/01/29/the-internet-of-things-has-a-growing-number-of-cyber-security-problems

Ray, P., (2016) "Internet of Robotic Things: Concept, Technologies, and Challenges," IEEE Access, vol. 4, pp. 9489-9500.

Reisig, W. (2013) "Understanding Petri Nets". Springer, 2013.

Stephenson, P. R. (2004). "A formal model for information risk analysis using colored Petri nets." in Proceedings of the Fifth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, Aarhus, Denmark, October 8-11, 2004, DAIMI PB - 570 / Kurt Jensen (Ed.), Oct, 2004, pp. 167–184.

Taipale, S., Vincent J., Sapio, B., Lugano, G., and Fortunati, L. (2015) " Social robots from a human perspective." Switzerland: Springer, 2015, pp. 11-15.

Tobe, F., (2017) "Robotics industry growing faster than expected - The Robot Report," The Robot Report, 2017. [Online]. Available: https://www.therobotreport.com/robotics-industry-growing-faster-than-expected/. [Accessed: 16- Sep- 2018].

Yang, Y., Dresher, S., and Kim, S., (2015) "Modeling, Simulation and Analysis of Autonomous Robot Navigation Algorithms using Fuzzy Petri Nets," International Conference on Computers And Their Applications.

PhD Research Papers

A Unified Forensics Analysis Approach to Digital Investigation

Ali Alshumrani[1,2], Nathan Clark[1,3] and Bogdan Ghita[1]
[1] Centre for Cyber Security, Communications and Network Research (CSCAN), University of Plymouth, UK
[2] Department of Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia
[3] Security Research Institute, Edith Cowan University, Western Australia

ali.alshumrani@plymouth.ac.uk
n.clarke@plymouth.ac.uk
bogdan.ghita@plymouth.ac.uk

Abstract: Digital forensics is now essential in addressing cybercrime and cyber-enabled crime but potentially it can have a role in almost every other type of crime. Given technology's continuous development and prevalence, the widespread adoption of technologies among society and the subsequent digital footprints that exist, the analysis of these technologies can help support investigations. The abundance of interconnected technologies and telecommunication platforms has significantly changed the nature of digital evidence. Subsequently, the nature and characteristics of digital forensic cases involve an enormous volume of data heterogeneity, scattered across multiple evidence sources, technologies, applications, and services. It is indisputable that the outspread and connections between existing technologies have raised the need to integrate, harmonise, unify and correlate evidence across data sources in an automated fashion. Unfortunately, the current state of the art in digital forensics leads to siloed approaches focussed upon specific technologies or support of a particular part of digital investigation. Due to this shortcoming, the digital investigator examines each data source independently, trawls through interconnected data across various sources, and often has to conduct data correlation manually, thus restricting the digital investigator's ability to answer high-level questions in a timely manner with a low cognitive load. Therefore, this research paper investigates the limitations of the current state of the art in the digital forensics discipline and categorises common investigation crimes with the necessary corresponding digital analyses to define the characteristics of the next-generation approach. Based on these observations, it discusses the future capabilities of the next-generation unified forensics analysis tool (U-FAT), with a workflow example that illustrates data unification, correlation and visualisation processes within the proposed method.

Keywords: Data Correlation, Data Heterogeneity, Digital Forensics, Digital Forensics Tools.

1. Introduction

The continuous development of technologies, network services and communications alongside the growing number of digital devices adopted and utilised by communities has enabled enormous data expansion. The adoption of modern technologies has also been accompanied by an increase in associated criminal activities, which recorded more than 5.1 billion data breaches in 2021 (Luke Irwin, 2022), presenting several challenges to digital forensics (Chabot, et al., 2015; Chikul, Bahsi and Maennel, 2021). With more than 90% of criminal activities leaving a digital footprint, digital forensics is arguably critical to deal with these crimes (Cellebrite, 2021). Evidence can now be extracted across many devices and delivered in various data formats such as structured, semi-structured and unstructured data (Reeve, 2013). However, the complexity of enormous data volumes and diverse data types, fragmented across different levels and spanning various digital sources, has changed the nature of digital evidence. The proliferation of file system platforms, smartphones, cloud applications, and IoT sensors has led to a constant increase in the volume of data and data heterogeneity at a high rate (Aggarwal and Davis, 2018). Unfortunately, to date, current methodologies and tools tend to be siloed around specific technologies or services. These isolated approaches to examination and analysis introduce significant challenges when aiming to cross-correlate evidence. For that reason, investigating multi-source complex digital information is challenging and requires substantial cognitive load and intensive manual labour (Adderley and Peterson, 2020).

The existing approaches and tools are not widely researched/developed with features to support multi-source correlation analysis. Instead, they primarily rely on experts' knowledge and cognitive ability to map the relationships between diverse data points (Mohammed, Clark and Li, 2018). It is evident that there is a crucial need to develop a robust, unified, automated method to process, integrate, analyse and correlate disparate data within one unified tool. One potential approach to discovering such activity is to cross-correlate evidence across data sources, reveal comprehensive data relationships, and identify inconsistent data or missing parts of evidence. The development of this approach could help to identify a complete set of evidence, reveal hidden data, detect inconsistencies, and identify anti-forensic activities in a timely manner. Therefore, this paper

discusses the future capabilities of a next-generation forensic analysis tool by presenting its ability to harmonise, correlate and visualise disparate data automatically, in a cognitively simple way, and in a forensically and timely fashion.

The remainder of this paper is structured as follows. Section 2 thoroughly evaluates the capabilities and limitations of current digital forensic tools. Section 3 discusses the current state of the art. Section 4 presents a novel approach to unifying digital forensics. Section 5 discusses the conclusions and directions of future research.

2. Evaluation of Digital Forensics Tools

A wide range of digital forensics tools and applications have been developed with features that aid digital forensics practitioners in acquiring, examining, analysing, and reporting digital evidence. This section provides an analysis of the current capabilities and limitations of forensic tools. The tools have been carefully selected based on several criteria, such as their diversity of features and functionalities, recency, and popularity among digital investigation practitioners. Finally, the chosen software is categorised and comprehensively evaluated based on their capabilities of the technical parameters associated with core digital forensics categories, as shown in Table 1.

Table 1: A comprehensive evaluation of digital forensics tools capabilities

	Digital Forensics Tools	Parameters	Acquisition & Examination Phase							Analytics Phase					
			Supported Data Source							Limited Data Sources			Automating Multisource Data Analysis		
			File System	Network	Smartphone	Social Media	IoT	Database	Cloud	Automated Analysis Supported	Basic Correlation Analytics	Basic Visualization	Anti-Forensics Detection	Advanced Correlation	Advanced Visualizati
File system	Autopsy		✓	×	✓	×	×	L	×	P&L	P&L	✓	×	×	×
	Detego Unified Forensics		✓	×	✓	✓	✓	N/A	✓	P&L	✓	✓	×	×	N/A
	EnCase		✓	×	✓	✓	✓	N/A	✓	P&L	×	✓	×	×	×
	Magnet Forensics AXIOM		✓	×	✓	✓	✓	N/A	✓	P&L	P&L	✓	×	×	N/A
	Nuix Investigate		✓	×	✓	✓	×	✓	✓	P&L	P&L	✓	×	×	N/A
	OS Forensics		✓	×	×	×	×	✓	×	×	×	✓	×	×	×
	Forensic Explorer		✓	×	×	×	×	×	×	L	×	✓	×	×	×
	FTK		✓	×	✓	✓	N/A	✓	✓	P&L	P&L	✓	×	×	×
	X-Ways Forensics		✓	×	×	×	×	✓	×	×	×	✓	×	×	×
Network	NetDetector Suite		×	✓	×	×	×	×	×	×	×	L	×	×	×
	Network Miner		×	✓	×	×	×	×	×	×	×	L	×	×	×
	Wireshark		×	✓	×	×	×	×	×	×	×	L	×	×	×
	Xplico		×	✓	×	×	×	✓	×	×	×	L	×	×	×
Smartphone	Belkasoft Evidence Center		✓	×	✓	✓	×	✓	✓	×	×	✓	×	×	N/A
	Cellebrite PathFinder		✓	×	✓	✓	✓	N/A	✓	P&L	P&L	✓	×	×	×
	Mobiledit Forensic Express		×	×	✓	✓	×	N/A	✓	P&L	×	✓	×	×	×
	Oxygen Forensic Detective		✓	×	✓	✓	✓	✓	✓	P&L	P&L	✓	×	×	N/A
	Paraben E3 Universal		✓	×	✓	✓	✓	✓	✓	×	×	✓	×	×	N/A
	XAMN Analyse		×	×	✓	✓	N/A	N/A	N/A	×	×	✓	×	×	×
Database	Sanderson Forensics		×	×	×	×	×	✓	×	×	×	×	×	×	×
	SysTools SQL Log Analyzer		×	×	×	×	×	✓	×	×	×	×	×	×	×

(✓): Addressed, (×): Not Addressed, (L): Limited Features, (P): Partially Automated, (N/A): Not Applicable.

It can be seen from Table 1 that these tools have been designed to operate within the confines of particular technologies, whether that be devices, such as computers and smartphones, or types of data, such as network data and databases. A wide range of tools exist, but many of them were developed for file-systems based evidence, for instance, Autopsy, EnCase, Nuix, FTK, and X-Ways forensic tools, with abilities to some extent to deal with other platforms, such as smartphones and internet applications. With a limited capability to perform advanced data analytics, they are limited to conducting data correlation within one single data source or basic data visualisation analysis. With regard to network investigation, a few network data analysis tools exist, such as Wireshark, NetworkMiner and Xplico, but with limited examination features to investigate network logs. Indeed, the network tools evidently lack incorporating any other types of evidence, as they were primarily

developed to monitor and troubleshoot network logs. Whereas smartphones tools, such as Belkasoft, Cellebrite, Modiledit, Oxygen, Paraben, and XAMN, were developed with a great feature to deal with digital resources apart from smartphones, such as file systems, web applications, IoT and cloud data. However, their functionalities mainly reside in fundamental investigation techniques, such as data acquisition and examination phases. Only partial attention is given to developing an advanced investigation method, such as evidence automation concerning multiple data sources.

The following points provide further detailed analysis of the capabilities of these tools against the current challenges in the digital forensics discipline:

- **Evidence Volume:** These investigation tools struggle to process and identify potential evidence among a large volume of data involving multiple evidence resources in a reasonable amount of time (Martinez-Mosquera, Navarrete and Lujan-Mora, 2020). Several experiments have been conducted to evaluate the efficiency of these tools, which show that several hours were taken to perform the acquisition and examination of data evidence (Horsman, 2019). For instance, an experiment aimed to identify items of interest using data signature analysis of data evidence contains 10 million files, and it concluded that tools such as Encase and Nuix required a long time to only identify potential data (Quick and Choo, 2018a). Technically, these experiments indicate that these tools perform poorly against the growing volume of data.
- **Data Heterogeneity:** The current tools are incapable of integrating a large volume of data heterogeneity; even though some of these tools have the ability, to some extent, to extract complex data, such as unstructured data, they still cannot address the issue of providing a complete understanding of evidence data without cognitive overload (Chabot, et al., 2015). Arguably, these tools do not support features that would integrate complex data of different formats automatically (Nordvik, Toolan and Axelsson, 2019). Moreover, the absence of data automation leads to the overloading of digital investigators when they attempt to handle and investigate complex data heterogeneity. Impressively, none of the tools evaluated could provide an investigator with the ability to query evidence across heterogeneous data in a unified and timely manner.
- **Data Correlation:** The developers of Autopsy, Detego, AXIOM, Nuix, and Oxygen state that these tools have data correlation capabilities. Interestingly, these tools can only perform partial data correlation within the same data sources. For instance, the features of the correlation engine module in Autopsy are limited to a few data correlation attributes, such as domain names, email addresses, and phone numbers. Moreover, it is not fully automated and requires forensic investigators to understand the logic of the case before the commencement of the data correlation process and, ultimately, to validate the results. Although the Magnet AXIOM tool utilises a built-in algorithm to determine and tag potential images with illegal content, such as weapons and drugs (Magnet, 2022), the investigator's intervention is still necessary to view and verify the results before analysing the content. Finally, the correlation analysis is limited solely to data file system metadata, with no attention given to recognising different metadata structures from other data sources.
- **Data Visualisation:** Conducting data visualisation analysis can enable examiners to visualise and uncover potential evidence effectively. The selected tools are limited in terms of data visualisation techniques that can function with a reasonable volume of data types. For instance, evidence visualisation in Autopsy is limited to timeline events and fails to combine timeline visualisation with geographical data, although it can extract Global Positioning System (GPS) data from multimedia files and applications. While AccessData reported that FTK can visualise data objects in different views in the same context, its design lacks an interactive navigation system, making it complex for an inexperienced investigator to use (AccessData, 2022). Moreover, the current approaches cannot visualise high-level system events along with the associated low-level traces in a user-friendly interface (Adderley and Peterson, 2020).

It is worth noting that these tools incorporate a wide range of forensic analysis techniques, such as facial recognition, explicit image detection, social network analysis, and data visualisation. Nevertheless, the primary usage of these tools remains in the early stages of the digital investigation process, as their strength resides in the data acquisition and examination phases. Furthermore, the evaluation also concludes that the selected tools cannot support an investigation of a high volume of heterogeneous data in a usable manner. Another limitation is the inability to perform advanced data analysis. For instance, none of these tools can correlate network traffic with another evidence source, such as file systems, or cross-correlate CCTV data with government databases to find data matches. Finally, none of these tools can perform advanced link analysis across all evidence resources within one user interface. Thus, the digital forensics field still lacks sufficient features to conduct a digital investigation in a forensically sound and timely manner.

3. Analysis of the Current State of the Art

To provide an appreciation of the current state of the art, an analysis was undertaken of academic publications related to data volume and heterogeneity, evidence correlation, and data visualisation analytics. Researchers have applied several methods, such as data ontology, machine learning and data graph, to prioritise potential artefacts, remove irrelevant and duplicate data, and correlate and visualise evidence.

Data ontology is beneficial for logically presenting data and relationships between classes of a specific domain. Brady, Overill and Keppens (2015) proposed a Digital Evidence Semantic Ontology (DESO) to organise and categorise relevant artefacts. DESO creates a repository of known artefact attributes, such as data type, location and reference class, to aid examiners in prospectively understanding data and selecting the relevant evidence. The experiment indicated that the DESO could only identify and represent data that remains in its identifiable data structure but fails to process unstructured data. Similarly, Chabot, et al. (2015) introduced Semantic Analysis of Digital Forensic Cases (SADFC). SADFC represents the ontological nature of corresponding data and uses SPARQL data query language to query evidence. During the evaluation, a real malware dataset was utilised, indicating that SADFC could extract related knowledge, perform data analysis, and execute data queries. However, it lacks the ability to automatically analyse and validate cases with a large volume of heterogeneous data. In the same context, Turnbull and Randhawa (2015) proposed Parallax Forensics (ParFor) framework that uses derive system events, such as login information and user events, including browser history and email events, and utilises them to reconstruct social network activities based data ontology mapping. However, this approach is limited to examining the abstraction of users' events and their attributes, whilst no attention has been given to low-level data, such as files and disks. Another framework that seeks to reduce evidence size is proposed by Quick and Choo (2018b). First, the approach seeks to create a logical evidence container and add an intelligence value to disparate evidence by identifying related entities via open-source information, such as social networking sites. The authors suggested using Open-Source Intelligence (OSINT) and developed Digital Forensic Intelligence (DFINT). The researchers utilised a real dataset to evaluate the proposed framework, which confirmed the possibility of semi-automatically scanning and combining data from disparate datasets across different data objects to some extent. However, further research is needed to enhance the system functionalities to reduce the investigation time.

Further ontological studies conducted by Amato, et al. (2020) and Amato, et al. (2019) aimed to employ a semantic-based methodology and Neuro-linguistic Programming (NLP) in their system to integrate and correlate evidence. First, this system utilises Resource Description Framework (RDF) assertions over the associated ontology to accomplish diversified data collection generated by investigation tools. It then uses its rule engine to integrate and correlate evidence data, and finally, a SPARQL module develops and evaluates users' queries against the SPARQL query engine. During the experiment, a collection of logging file events was used, which indicated its ability to detect potential events. Still, it functions to a limited type of data and requires manual intervention to help annotate events. With the same objectives, Chikul, Bahsi and Maennel (2021) developed a ForensicFlow system based on a semantic web to allow knowledge integration and semantically query data relationships. This ontology consists of data sources, data extraction modules, data analysis layers, data knowledge layers, and finally, event-based and forensic artefacts that aim to standardise and define different data types. While evaluating the capabilities of the proposed approach, the researchers simulated a ransomware dataset containing file system data and memory dump and applied the proposed system to the simulated data. The experiment result stated that the system is valuable in revealing and querying evidence data from one data source, but it fails to perform data validation.

Different studies sought to identify data correlation based on metadata analysis. Raghavan and Raghavan (2013) implemented the AssocGEN engine to analyse and determine associations among evidence data. The experiment result indicated that AssocGEN could not support all metadata attributes of file systems applications, making it unsuitable for dealing with the varieties of file systems data that existed today. It also indicates that the data correlation aspect still needs improvement to automate the identification of corresponding data without requiring investigator interventions. Likewise, Mohammed, Clark and Li, (2018), introduced a similar concept for determining artefact associations by developing an algorithm to merge different datasets through data characterisation and harmonisation. The characterisation process identifies the nature of the data through the metadata, and the harmonisation process combines the data. The evaluation shows that this algorithm performs optimally in some respects but cannot integrate all binary data fields, making the harmonisation process less accurate. Finally, these two approaches are limited to some data types, making them less robust and generalised for addressing current crime categories.

Noel, et al. (2016) proposed an alternative approach named CyGraph based on the Neo4j data graph technique. The CyGraph aims to unify, correlate, and visualise network events. First, the system collects and stores network events, security events, and the associated vulnerabilities. It then utilises its built-in model to predict possible attack paths and critical vulnerabilities across cyber entities. To some extent, this approach can aid examiners in executing data queries and expressing graph patterns visually. However, it only covers the predefined data in the data model and does not allow the investigator to manipulate data relationships interactively. In a further approach, Aggarwal and Davis (2018) attempted to standardise and integrate evidence data by converting relational and RDF models into data graphs by leveraging the (Bouhali and Laurent (2015) method. The approach first transforms RDF models by exporting databases into a CSV file format to allow the user to select the schema and transform it into a Neo4j data graph. Then, the Neo4j browser is used to graphically explore and interact with the transformed schemas. However, this approach is incapable of capturing schema conflicts, and it also fails to identify data subset relationships.

In a further effort, Okolica 2017) developed the Temporal Event Abstraction and Reconstruction (TEAR) tool, and employed a machine learning algorithm to aid in confirming the identification of data patterns. Technically, this algorithm initiates a hierarchy of system events and applies a pattern-matching technique to highlight a high-level event with low-level activities. Similarly, Schelkoph, Peterson and Okolica (2019) leveraged the advantages of TEAR and incorporated it to develop the Property Graph Event Reconstruction (PGER) system. The PGER system aims to normalise and identify data correlation by adapting a native graph property of a zero-index traversal to help discover adjacent nodes of related system events, such as web history and download data. However, both TEAR and PGER methods need to be evaluated on a larger scale of data types and a greater data volume. Later, Adderley and Peterson (2020) leveraged the PGER system and introduced the Temporal Analysis Integration Management Application (TAIMA). TAIMA aims to visualise graphically and reconstruct evidence events based on a timeline analysis mechanism. First, TAIMA recognises timestamp events as a research parameter within file system data attributes, such as event logs, registry logs, and link files. Then, it displays the identified high-level events with their associations on a single screen. The researchers utilised a simulated case study to evaluate the effectiveness of TAIMA; the results highlighted that TAIMA could conduct timeline analysis and filter the number of system events based on timeline analysis. However, its functions are limited to timeline analysis and only recognise executable files and web history data. Finally, the analysis presented above is summarised in Table 2.

Table 2: Summary of the literature review

N	Author(s)	Context	Objective(s)	Method	System Name	Mode	Limitation
1	(Brady et al., 2015)	DV	Categorisation of potential data.	Ontology	DESO	Case Study	• Partially automated. • Applicability in complex data.
2	(Chabot et al., 2015)	DV & DH	Automatically represent evidence events.	Ontology	SADFC	Simulated	• Lacks to validate data automatically. • Scalability.
3	(Chikul et al., 2021)	DH & DC	Automatically extract and reconstruct disparate data.	Ontology	Forensic Flow	Simulated	• Limited to specific data types. • Lacks to validate data queries. • Scalability.
4	(Turnbull and Randhawa, 2015)	DH	Provide a unified representation of multiple data sources.	Ontology	ParFor	Case Study	• Data validation. • Limitation in data correlation. • Applicability in real data.
5	(Amato et al., 2020, 2019)	DH & DC	Integrate & correlate evidence data.	Ontology & NLP	NA	Simulated	• Partially automated. • Limitation in data types.
6	(Quick and Choo, 2018)	DV & DH	Identify the related entity.	Framework	DFINT & OSIT	Real Data	• Limitation in data types. • Investigation time.
7	(Raghavan and Raghavan, 2013)	DH & DC	Identify data association	Engine	AssocGE-N	Real Data	• Limitation in data types. • Partially automated.
8	(Mohammed et al., 2018)	DH	Data characterisation & harmonisation.	Algorithm	NA	Real Data	• Limited data sources. • Lack to merge all binary data.
9	(Noel et al., 2016)	DC & DVis	Unification model to correlate and visualise network events.	Data Graph	CyGraph	Case Study	• Applicability in complex data. • Scalability & usability.
10	(Aggarwal and Davis, 2018)	DH	Transform relational and RDF models into data graph properties.	Data Graph	NA	Case Study	• Lacks to capture all data. • Applicability in real data.
11	(Okolica, 2017)	DH	Correlate evidence data.	Data Graph & ML	TEAR	Real Data	• Limited data types. • Applicability in complex data.
12	(Schelkoph et al., 2019)	DC	Normalise and correlate evidence data.	Data Graph & ML	PGER	Real Data	• Limited data types. • Applicability in complex data.
13	(Adderley and Peterson, 2020)	DC & DVis	Graphically correlate & visualise evidence data.	Data Graph	TAIMA	Case Study	• Limited evidence events. • Timeline analysis only. • Applicability in real data.

DV: Data Volume, DH: Data Heterogeneity, DC: Data Correlation, and DVis: Data Visualisation

Researchers have contributed to overcoming challenges in digital forensics by proposing several methods and techniques based on technologies such as data ontology, machine learning, and data graphs. Unfortunately, most of these methods attempt to articulate a specific obstacle or focus on a particular data type. Thus, many limitations remain, indicating the need for further research to fill the identified gaps, enhance investigation productivity, and overcome existing challenges. For instance, a comprehensive research should include data unification and data automatically to merge disparate data and provide usability to ensure that examiners can use the solutions in a way that requires little effort in identifying and correlating potential evidence across various data within one graphical interface. Therefore, this study emphasises the necessity of establishing a unified forensics analysis solution to integrate and unify the ubiquitous nature of data heterogeneity and identify the relationships between all digital events and user activities across different sources of evidence.

4. Next-Generation U-FAT

To facilitate a comprehensive understanding of the functionalities and features required of a next-generation *Unified Forensics Analysis Tool* (U-FAT), a list of common crimes was compiled, categorised and linked with the necessary corresponding digital analyses to address investigation questions and meet investigation objectives. Having defined these core requirements, U-FAT is proposed and discussed.

4.1 Crime Classification

Current criminal investigation cases have come to be involved in interconnected data from across different technologies, requiring advanced digital forensics analysis techniques that must function alongside each other to finalise an investigation in a timely fashion. Figure 1 summarises some of the most common crimes, associating them with examples of digital evidence sources. It also highlights some types of forensic analysis that could be employed to conduct criminal investigations and address the investigation objectives.

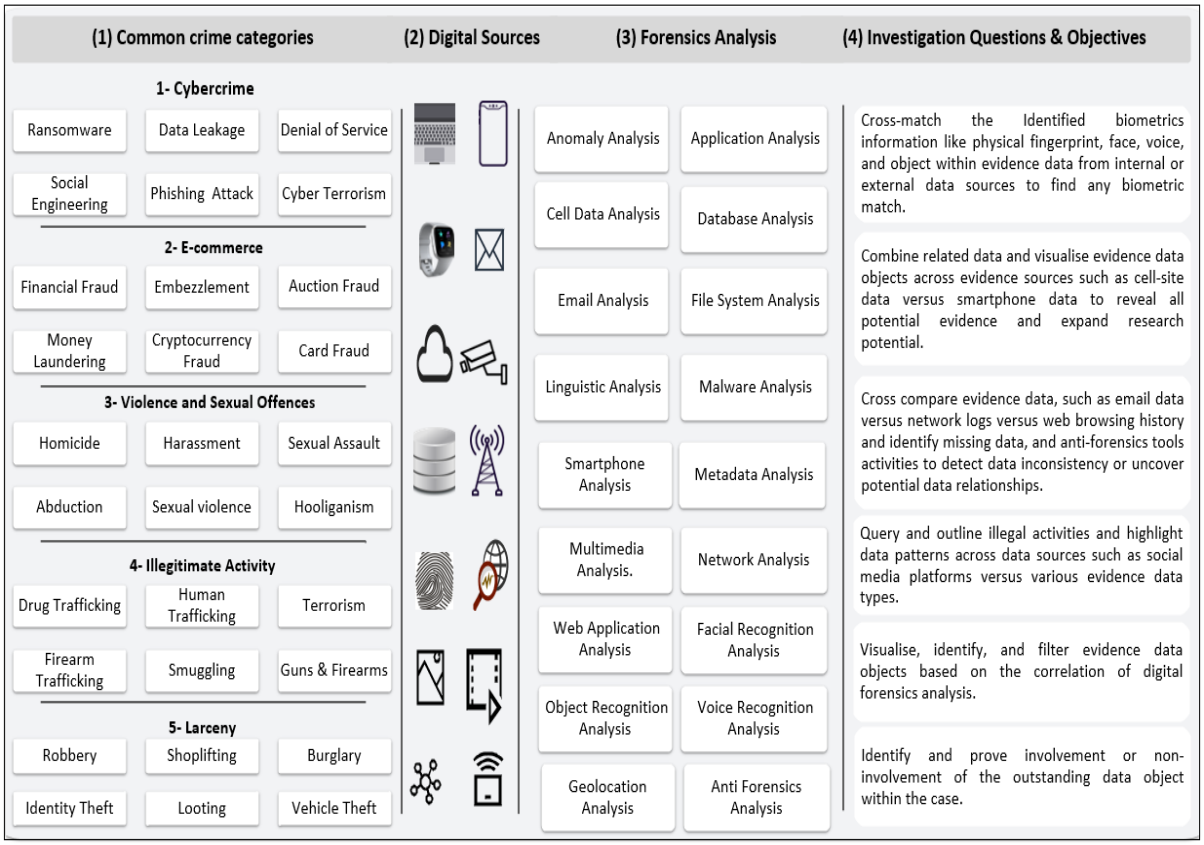


Figure 1: Classification of crimes and forensics analyses

It is worth noting that in Figure 1, each crime might involve a variety of data sources containing different data formats, which in turn require several types of forensic analysis to conduct a proper digital investigation. Two examples are provided below as an illustration:

- **Cybercrime:** Investigating cyber hacktivist activities, such as those involving ransomware, denial-of-service attacks, social engineering, and online scams, can be challenging, as these crimes involve multiple connected computers, cloud system resources, applications, and network devices. Therefore, several forensic examinations are required based on each data source, for example: file system, network, malware, database, and application analyses.
- **Violence, illegitimate activity, and larceny crimes:** Although these types of offences might not be categorised as digital crimes, still, digital data can be employed to support an investigation. Evidence could come from digital devices, such as laptops and smartphones, or external data, including public internet applications, CCTV, and call data records. Whilst investigating such data, digital investigators will perform different forensic analyses, such as: application, multimedia, linguistic, biometric, social network, cell data records, databases, and social network assessments to identify potential evidence and correlate the results to establish appropriate answers.

In summary, Figure 1, highlighted the complexity of current digital investigation cases, the limitations of digital tool capabilities outlined in Table 1, and the existing gaps in digital forensics methods pinpointed in Table 2 have formed the main drive underpinning in this research.

4.2 System Workflow

The proposed approach comprises a unified system capable of forensically performing end-to-end the investigation of disparate data in a time-efficient, cognitively simple, and conclusive manner. A holistic workflow of the proposed U-FAT system is illustrated in Figure 2. It shows various featured functions to perform data analytics side by side, for instance: mapping evidence components, detecting anomalies and data inconsistencies, revealing data query matching, discovering data patterns, and permitting cross-platform analysis of complex, interconnected data within a single unified graphical interface.

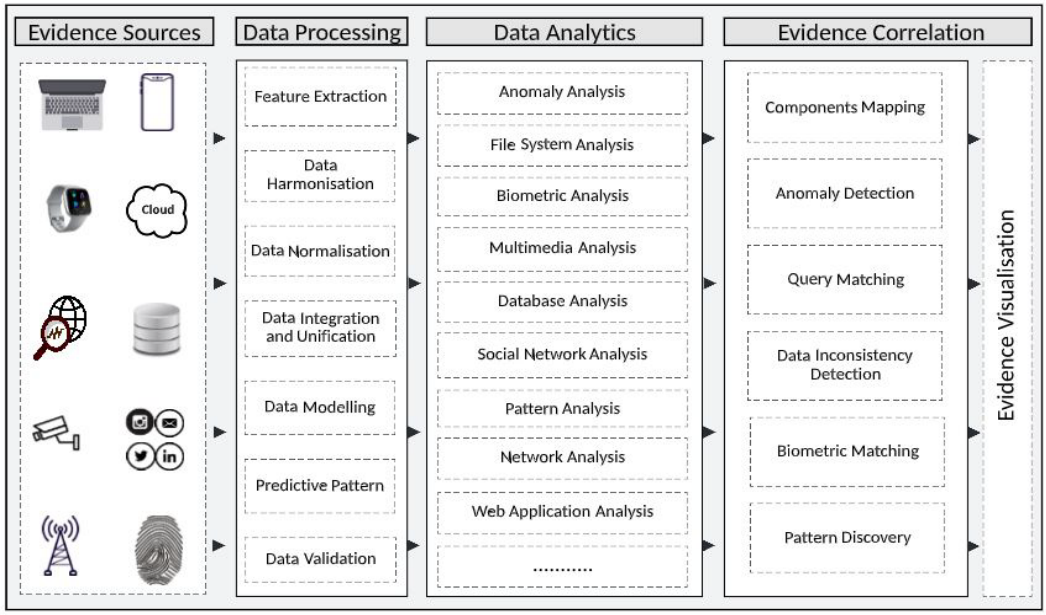


Figure 2: Holistic illustration of the U-FAT process

The initial functionalities of the system aim to define and extract disparate data depending on different data filtering and extraction mechanisms. Then, various data examination techniques are applied in the pre-processing stage to eliminate duplicate or irrelevant data, identify potential data and pass it to the data processing phase. In the data processing phase, the system applies data normalisation, harmonisation, and unification techniques, to interpret complex data into a standard and appropriate normalised data format, grouping and unifying them based on their attributes and characteristics. Afterwards, different evidence analytics methods will be undertaken to detect data anomalies, patterns, biometrics, and many more based on the requirements of the investigation.

To technically illustrate the proposed approach, Figure 3 presents a workflow of the data unification and correlation processes in a case involving the need to analyse evidence data, such as: file system, email, network logs, web applications, and social media, as this could be the case in many investigation crimes.

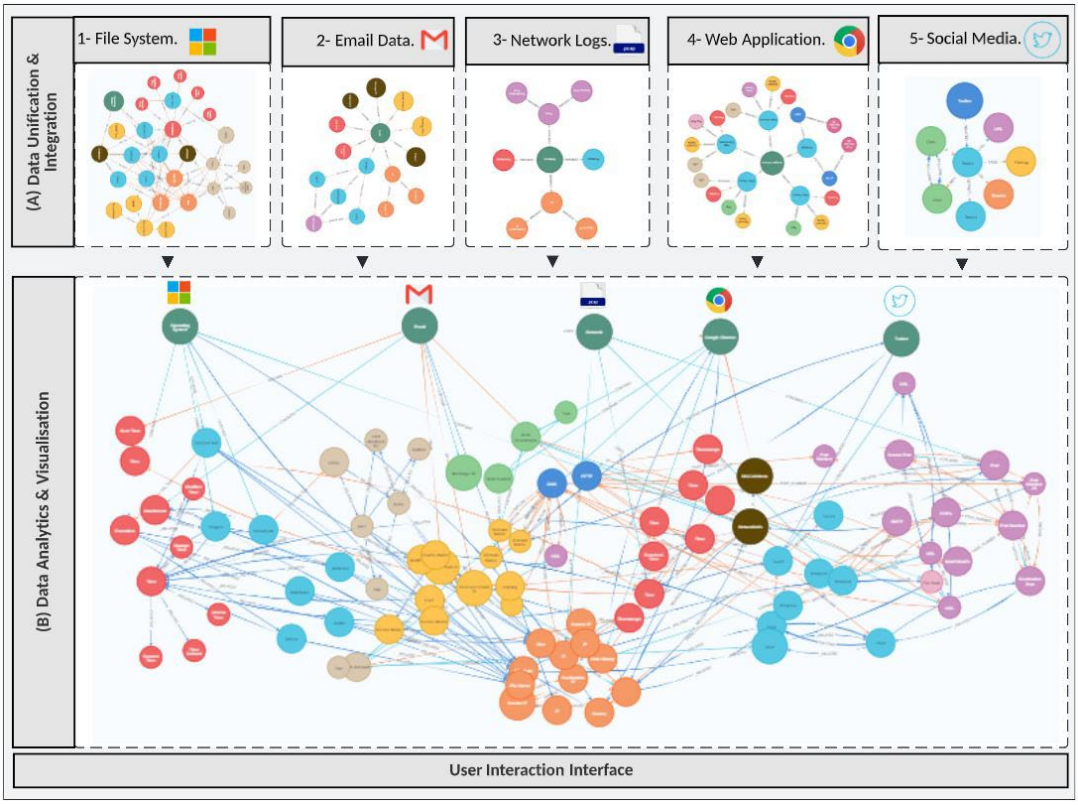


Figure 3: Workflow example of data unification, correlation, and visualisation in Neo4j data model.

The main benefit of the data unification technique is to enable advanced forensic analysis across different data platforms. Figure 3 illustrates the core functional components of the U-FAT, where the initial process aims to identify different data sources, extract and categorise data based on each data attribute, and then utilise its built-in mechanisms to unify and correlate data in a unified view of complete evidence. Figure 4 presents a technical example of how these processes have enabled advanced forensic analysis.

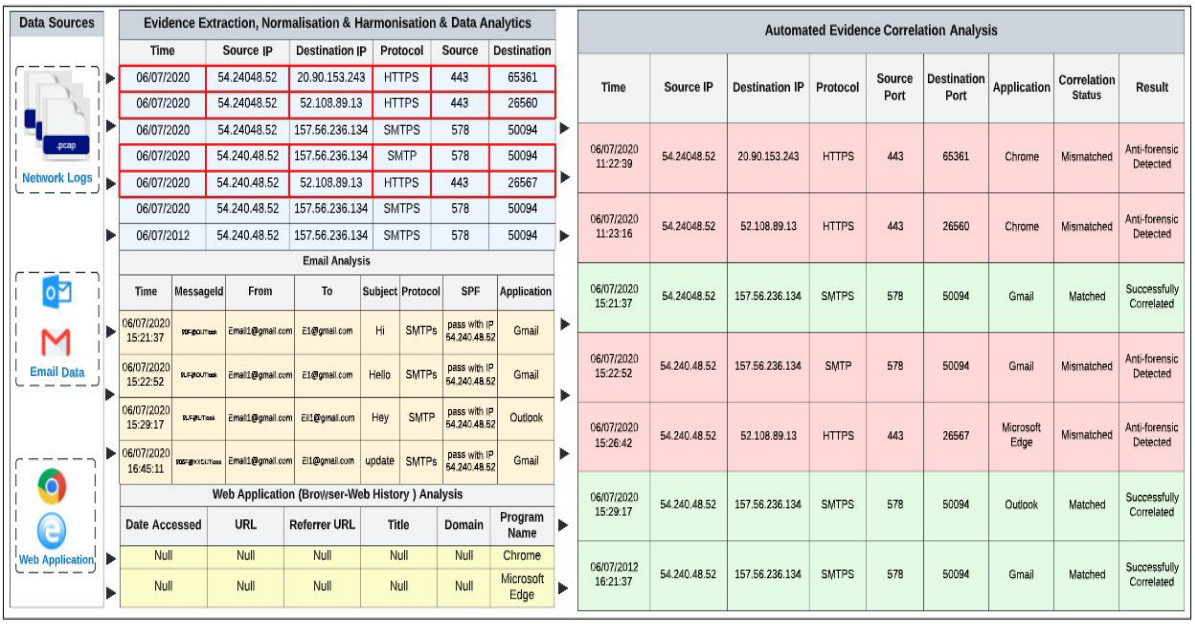


Figure 4: Workflow example of data harmonisation and correlation processes.

In this example, the proposed approach utilised metadata attributes of evidence data, including network logs and email data from applications such as Outlook or Gmail, and historical data of web applications such as Google Chrome or Microsoft Edge, to automatically harmonise and unify in a standardised format. This approach would enable advanced data analytics to function harmoniously against harmonised data automatically; for instance, mapping artefacts against a single timeline, data correlation analysis and detecting the presence of anti-forensics techniques. Figure 4 indicates that it is possible to identify when a suspect might have been using data-hiding techniques, such as private web browsing or deleting emails from the client machine. Web histories from a suspect machine can be cross-correlated with network traffic to identify them, and the client email records can be compared against network and server-side records, therefore, contributing to addressing the current challenges in digital forensics and reducing and eliminating the investigation time and manual cross-correlation.

In the U-FAT approach, various analytics modules are integrated and function side by side to perform advanced evidence analytics. The analytics automation process provides the capability to harmonise, query, correlate and analyse evidence across evidence data. Notably, the correlation analytics stage is considered the cornerstone of the U-FAT approach, enabling digital investigators to work interactively to query, review and edit data relationships across a complete evidence case. Having the ability to retrieve, correlate and graphically visualise potential data within one tool could provide more insight into the output of the analytics process. Empowering forensic investigators with the proposed approach can considerably shorten the investigation time and reduce the cognitive for more significant insights to be gained through different visualisation techniques in respect of heterogeneous data with different views.

5. Conclusion

Unifying heterogeneous evidence and incorporating digital forensics analysis has become increasingly important in digital forensics investigations. Current approaches offer a range of forensics analyses and semi-automated data techniques. However, the methods discussed have yet to be able to process, integrate and correlate a vast amount of data heterogeneity in an automated and harmonised fashion. Therefore, it is essential to integrate and unify multiple data sources and process advanced analytics in one intuitive interface to overcome the limitations and complexity of the current frameworks of evidence investigation. Furthermore, an intelligent query feature is proposed to support the flexibility of enabling data interrogation functions across technologies and services. Combining these features with an automated mechanism to understand, correlate and visualise data intelligently will aid digital investigators in addressing questions in a more timely and cognitively efficient manner. Future research will focus on developing a common forensic information interchange language to permit cross-platform disparate case analysis and on developing advanced forensic analysis capability.

References

AccessData, 2022. Zero in on evidence faster—recognized around the world as the standard in computer forensics software, [online]. Available at: <https://accessdata.com/assets/pdfs/LIT-FTK-7.4.2.pdf> [Accessed 10 November 2022].

Adderley, N. and Peterson, G., 2020. Interactive temporal digital forensic event analysis. In: IFIP *Advances in information and communication technology*. pp. 39–55. https://doi.org/10.1007/978-3-030-56223-6_3

Aggarwal, D. and Davis, K. C., 2018. Employing graph databases as a standardization model for addressing heterogeneity and integration. In: Rubin, S.H., Bouabana-Tebibel, T. eds., *Advances in intelligent systems and computing*. Springer International Publishing, Cham, pp. 109–138. https://doi.org/10.1007/978-3-319-56157-8_6

Amato, F., Castiglione, A., Cozzolino, G. and Narducci, F., 2020. A semantic-based methodology for digital forensics analysis. *Journal of Parallel and Distributed Computing,* 138, pp.172–177. https://doi.org/10.1016/j.jpdc.2019.12.017

Amato, F., Cozzolino, G., Moscato, V. and Moscato, F., 2019. Analyse digital forensic evidences through a semantic-based methodology and NLP techniques. *Future Generation Computer Systems,* 98, pp.297–307. https://doi.org/10.1016/j.future.2019.02.040

Bouhali, R. and Laurent, A., 2015. Exploiting RDF open data using NoSQL graph databases. In: IFIP *Advances in Information and Communication Technology*. New York: Springer. pp. 177–190. https://doi.org/10.1007/978-3-319-23868-5_13

Brady, O., Overill, R. and Keppens, J., 2015. DESO: Addressing volume and variety in large-scale criminal cases. *Digital Investigation*, 15, pp.72–82. https://doi.org/10.1016/j.diin.2015.10.002

Cellebrite, 2021. Digital Intelligence Benchmark Report, [online]. Available at: <https://cellebrite.com/en/digital-intelligence-benchmark-report-2021> [Accessed 10 November 2022].

Chabot, Y., Bertaux, A., Nicolle, C. and Kechadi, T., 2015. An ontology-based approach for the reconstruction and analysis of digital incidents timelines. *Digital Investigation*, 15, pp.83–100. https://doi.org/10.1016/j.diin.2015.07.005

Chikul, P., Bahsi, H. and Maennel, O., 2021. An ontology engineering case study for advanced digital forensic analysis. In: *Lecture notes in computer science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 67–74. https://doi.org/10.1007/978-3-030-78428-7_6

Horsman, G., 2019. Tool testing and reliability issues in the field of digital forensics. *Digital Investigation,* 28, pp.163–175. https://doi.org/10.1016/j.diin.2019.01.009

Luke Irwin, 2022. Free Infographic: List of data breaches in 2021 [online]. IT Gov. Eur. Blog. Available at: <https://www.itgovernance.co.uk/infographics/list-of-data-breaches-in-2021> [Accessed 11 January 2022].

Magnet, 2022. Magnet IEF - Artifact-First Investigation [online]. Available at: <https://www.magnetforensics.com/products/magnet-ief/> [Accessed 22 September 2022).

Martinez-Mosquera, D., Navarrete, R. and Lujan-Mora, S., 2020. Modeling and management big data in databases—a systematic literature review. *Sustainability,* 12, p.634. https://doi.org/10.3390/su12020634

Mohammed, H., Clarke, N. and Li, F., 2018. Automating the harmonisation of heterogeneous data in digital forensics. In: *European conference on information warfare and security*, ECCWS. pp. 299–306.

Noel, S., Harley, E., Tam, K. H., Limiero, M. and Share, M., 2016. CyGraph: Graph-based analytics and visualization for cybersecurity. In: *Handbook of statistics*. Elsevier B.V., pp. 117–167. https://doi.org/10.1016/bs.host.2016.07.001

Nordvik, R., Toolan, F. and Axelsson, S., 2019. Using the object ID index as an investigative approach for NTFS file systems. *Digital Investigation,* 28, pp.S30–S39. https://doi.org/10.1016/j.diin.2019.01.013

Okolica, J.S., 2017. Temporal event abstraction and reconstruction. Air Force Institute of Technology Wright-Patterson AFB United States

Quick, D. and Choo, K.-K.R., 2018a. IoT device forensics and data reduction. *IEEE Access 6*, pp.47566–47574. https://doi.org/10.1109/ACCESS.2018.2867466

Quick, D. and Choo, K.-K.R., 2018b. Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT+OSINT): A timely and cohesive mix. *Future Generation Computing Systems,* 78, pp.558–567. https://doi.org/10.1016/j.future.2016.12.032

Raghavan, Sriram and Raghavan, S. V., 2013. AssocGEN: Engine for analyzing metadata based associations in digital evidence. In: 2013 *8th International workshop on systematic approaches to digital forensics engineering (SADFE*). IEEE, pp. 1–8. https://doi.org/10.1109/SADFE.2013.6911541

Reeve, A., 2013. *Managing data in motion.* Elsevier. https://doi.org/10.1016/C2011-0-07758-X

Schelkoph, D. J., Peterson, G.L. and Okolica, J.S., 2019. Digital forensics event graph reconstruction. In: *Lecture notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Springer Verlag, pp. 185–203. https://doi.org/10.1007/978-3-030-05487-8_10

Turnbull, B. and Randhawa, S., 2015. Automated event and social network extraction from digital evidence sources with ontological mapping. *Digital Investigation*, 13, pp.94–106. https://doi.org/10.1016/j.diin.2015.04.004

Offensive Cyberspace Operations for Cyber Security

Gazmend Huskaj[1]
[1]Geneva Centre for Security Policy, Geneva, Switzerland

g.huskaj@gcsp.ch

Abstract: This work-in-progress research product covers Offensive Cyberspace Operations for Cyber Security or "Offensive Defense" for Cyber Security. Offensive cyberspace operations are shrouded in secrecy. From an intelligence perspective, this makes sense because of their development since Operation Desert Storm in 1991. The phenomenon, dubbed "Information Warfare," and to the professionals' surprise, they could remotely turn off an Iraqi power substation. However, the implication of remotely turning off the power substation was not only to cut off the power source to an Iraqi military headquarters, but it also meant cutting off the power to a nearby hospital, risking the lives of injured Iraqi soldiers protected by the Geneva Conventions. Since the 2000s and onwards, and with the US military recognizing cyberspace as a war-fighting domain, establishing United States Cyber Command (USCYBERCOM) may be a milestone. Thus, researchers have put much thought into cyberspace operations (offensive, defensive), such as doctrine, organizations, training, materials, leadership and education, personnel, facilities, and policy. One phenomenon, dubbed "defending forward," was coined in the 2018 US Department of Defense Cyber Strategy. The idea is simple: take the fight to the adversary. Other terms include "hunt-forward operations" and "offensive defense." Therefore, what is "Offensive Defense" for cyber security, and why now?

Keywords: Cyber Security, Information Warfare; Offensive Cyberspace Operations, Offensive Defense.

[1]*The views expressed in this paper are only those of the author, and do not represent the Geneva Centre for Security Policy (GCSP), or any other party.*

1. Introduction - Why Offensive Cyberspace Operations for Cyber Security?

Offensive cyberspace operations for Cyber Security is a relatively new concept. The idea is to strike a threat actor's targets through offensive methods to impose cost (Fischerkeller & Harknett, 2017) on the adversary to alter the threat actor's decision-making. Inferring costs could be by attacking critical infrastructure necessary for the threat actor to conduct their operations. By taking the fight to the adversary, it may be possible to make them switch stance: from an offensive stance to a defensive stance, for example to secure their systems.

Researchers have described the issue with deterrence and argued that the concept applied in the physical domains (air, land, sea, and space) does not work in the same matter in cyberspace. The research question is, therefore: What are offensive cyberspace operations for cyber security, and why now? Offensive cyberspace operations are defined as:

"a sequence of planned actions executed by an organised group of people with a defined purpose in and through hardware and software which are used to create, process, store, retrieve and disseminate information in different types of interconnected networks that build a large, global network, built and used by people. The offensive aspect includes methods to affect an adversary's target system's confidentiality, integrity and/or availability. These targets include 'weapon systems, C2 processes, logistics nodes, high-value targets'"

(Huskaj & Wilson, 2020, p. 513)

This article begins with a review of the scientific literature identifying the concepts of persistent engagement, defending forward, and offensive cyberspace operations for cyber security. The review output is a conceptual framework of elements needed for offensive cyberspace operations for cyber security, depicted in Table 1. The methods section is presented in 3, while the results in 4 and the discussion in 5.

2. Reviewing the Literature

In 2017, Fischerkeller & Harknett wrote an essay arguing that protecting or advancing national interests could not merely rely on deterrence. Second, shaping behavior in cyberspace requires conducting "active cyber operations" (p. 382). Their rationale is that because many cyber attacks have taken place, the deterrence strategy within the assumption that threat actors would restrain from conducting attacks is not working. One of the main reasons behind this rationale is the nature of the cyber domain. The cyber domain is an operational environment with information systems interconnected in networks of networks. Because of this, threat actors use this low-entry high-impact domain to influence, a domain that cuts through all the other domains of national power. Therefore, connected entities are always in contact with each other. Instead of having this deterrence strategy, the idea is to have a forward-leaning approach with persistent engagement. One year later, James A.

Lewis (Lewis, 2018) presented his perspectives in a book chapter about risk, resilience, and retaliation: American perspectives on international cyber security.

Lewis (2018) notes that because information systems are always vulnerable, more than a defensive strategy against persistent threat actors is required. He continues to note that states that are "not constrained by the etiquette and rules of Western law" (p. 252) will exercise cyber power by combining it with "conventional military and intelligence assets" (p. 252). Just like Fischerkiller & Harknett, Lewis (2018) argues that a defensive posture is insufficient, and the notion of "imposition of consequences" must be considered. In other words, it is about imposing costs on a threat actor's behavior and actions in cyberspace. Imposing costs is achieved by "persistent engagement" and "collective deterrence": like-minded states respond to malicious actions in cyberspace that violate UN-developed norms of responsible state behavior. Lewis (2018) continues to discuss potential risks with a "persistent engagement"-strategy, such as the risk of escalation, and that policy for this should be in place. He concludes that U.S. cybersecurity policy is more confrontative with a greater risk of warfare. However, not acting and remaining a victim is not that way forward.

Jason Healey (Healey, 2019) provides the background story of the policy direction of persistent engagement and the possible actions the new strategy might have. he presents the topics of strategic stability, superiority, and forward defense. Regarding strategic stability Healey (2019) states that the Internet is a complex system, and small input can generate unintended effects. The continuation of persistent engagement may create a far more insecure system than today which would impact trust. Furthermore, Healey (2019) notes that forward defense in cyberspace could be one of those problems that do not provide more stability but can make things worse, and the process of understanding this new environment is as with nuclear weapons; it is going to take decades because this is not only persistent it is permanent. He continues to present the importance of rhetorics concerning, for example, the difference between offensive cyberspace operations (OCO) and defensive cyber operations-response actions (DCO-RA), where the main difference between OCO and DCO-RA lies in the fact that DCO-RA missions may include actions that rise to the level of use of force with physical damage or destruction of enemy systems. At the same time, OCO is "intended to project power in and through foreign cyberspace [to] target adversary cyberspace functions or create first-order effects in cyberspace to initiate carefully controlled cascading effect" (Healey, 2019, p.2). Finally, he presents the changes to policy, from Presidential Policy Directive 20 (PPD-20) to National Security Policy Memorandum-13 (NSPM-13), delegated authority to the dual-hat chief of the National Security Agency (NSA) and US CYBERCOM for OCO: a shift from the Obama administration to the Trump administration.

In 2019, Healey & Jenkins presented a policy framework to assess whether the U.S. policy in persistent engagement is "suppressing or encouraging attacks." The idea here is like the concept of "battle damage assessment" but on the policy level.

Kosseff (2019) looks at defending forward (persistent engagement) from an international law perspective. More specifically, what opportunities and limits does international law provide? He concludes that international "provides the United States with significant leeway to position itself to degrade adversaries' cyber operations, collect information about cyber threats, and discourage other states from acting against the United States in cyberspace" (p. 1).

Sebekin (2020) analyses the U.S. policy on persistent engagement. He states that while there is an opportunity to shape acceptable behavior in cyberspace, so are the related risks with possible escalation. He concludes that "the prospects of this strategy are ambiguous" (p. 96), and only through the "accumulation of empirical experience in the application of this strategy" (p. 96) will the advantages and risks become apparent. The author argues that state relationships between Russia and the United States require an understanding of "U.S. approaches to cyber security and key strategic changes to them" (p. 96) to help shape Russian cyber security policy and build a "constructive dialogue with Washington on cyber security issues" (p. 96).

Table 1: Summarizes the concepts from the review discussion.

Concept	Definition	Key references
Deterrence	Through the threat of attack, or the conduct of attack, alter the adversary's calculus by enforcing costs	(Fischerkeller & Harknett , 2017; Lewis, 2018; Healey, 2019)
Persistent engagement	Using offensive methods in the adversary's infrastructure to shift their posture from offensive to defensive	(Fischerkeller & Harknett , 2017; Lewis, 2018; Healey, 2019)
Policy	Necessary policy in place that enables a cyber unit to conduct offensive cyberspace operations for cyber security	(Fischerkeller & Harknett , 2017; Lewis, 2018; Healey, 2019)
Frameworks	Frameworks to measure if policy on persistent engagement is suppressing or encouraging attacks	(Healey & Jenkins, 2019)

Concept	Definition	Key references
International law	What does international law state in relation to offensive cyberspace operations below the threshold of an armed attack	(Kosseff, 2019)
Diplomatic Dialogue	By understanding other states' policies help shape own polities for constructive dialogue	(Sebekin, 2020)

3. Methods and Materials

The research approach to this article was to search the Scopus database, the world's largest academic database of abstracts. Using the keywords {persistent engagement} reveals 74 document results. The results included many other research fields, such as Agriculture, Biology, Medicine, and Neuroscience. A manual review of the results revealed 14 research articles on Persistent Engagement. Manual screening of the 14 research articles presents six articles for review. The next step was to review open sources for empirical evidence on persistent engagement. The purpose was to identify cases for review to answer the research question and, more precisely, the second part: why now?

4. Results

Research articles on persistent engagement, defending forward, and offensive cyberspace operations for cyber security have increased over time. In 2017, one article presented why deterrence in cyberspace does not work as it does in the physical domains. In 2018, only one article provided the risks and opportunities of this new policy strategy. In 2019, researchers published three research articles reviewing historical developments in the field and providing a rationale for why deterrence in cyberspace was not working with different cyber operations concepts and a policy shift from one administration to another. The same year, another researcher wrote about how international law could allow persistent engagement. According to that author, international law provides significant leeway. Finally, in 2020 a Russian researcher provided an article to analyze US policy on persistent engagement and why this kind of research is essential for understanding another nation's cyber security policies and to help shape Dialogue between the two capitals.

Deterrence is a strategy applied in the physical domains. The early thinking was about applying deterrence in cyberspace. However, as the review provides, strategists and researchers considered it not applicable in cyberspace. Because the operational environment, information systems in networks of networks, is unique in many causes compared to the traditional domains. Some examples include proximity, speed, and impact. States, while geographically separated, in cyberspace are only milliseconds away. While combat aircraft can fly at 3.2 Mach, information systems can send instructions at speeds close to light speed, and while a navy ship can fire cruise missiles, the impact can be devastating. In cyberspace, it is possible to alter the instructions of a system to create all the effects as in physical space. However, it is also possible to generate disruptive effects through operations below the threshold of armed attacks.

A policy that dictates the use of offensive cyberspace operations within an overarching national strategy for cyber security is essential. Policy-makers must understand all the technical underpinnings of offensive cyberspace operations and the operational environment. However, they understand the impact and consequences of failed operations. It is ultimately the policy level that takes the political risk. According to open sources, when President Bush was in a hand-over/take-over with President Obama, President Bush asked President Obama to continue with two programs: the "drone program" and a program on offensive cyberspace operations. The Obama administration understood the importance of OCO. However, they likely did not feel comfortable delegating their use to USCYBERCOM. Reviewing the open-source literature indicates that USCYBERCOM became more mature as they conducted OCO. With the shift from the Obama administration to the Trump administration, they delegated decision-making to conduct OCO to the head of the NSA and USCYBEROM. It is likely an indication of an administration willing to take more political risk and that USCYBERCOM was becoming more mature.

International law is necessary and tries to govern the behavior of states. However, international law does not prohibit the work of intelligence agencies. Therefore, cyberspace operations for intelligence, surveillance, and reconnaissance (ISR) are not prohibited. Cyber-ISR operations can occur over time and, like traditional intelligence operations, do not violate international law.

Frameworks can support the planning, preparation, and execution of cyberspace operations. Frameworks can also support battle damage assessment and the measurement of policies for OCO and Cyber-ISR. It is essential to have frameworks to measure success in the mentioned areas. Hence, it is crucial to continuously feed new knowledge into the frameworks through a cybernetic feedback loop and update them accordingly.

Diplomatic Dialogue ensures that communication between states' occurs at all levels, from peace to crisis and un-peace. States have realized they can avoid mentioning war because they can achieve impact through cyberspace and hybrid means without declaring war. Declaring war is problematic as a lot of legal instruments are triggered.

5. Discussion

The answer to the research question, *What are offensive cyberspace operations for cyber security, and why now?* is that more empirical evidence exists on applying OCO for deterrence and cyber security. Cases like Operation Glowing Symphony presents how USCYBERCOM, with US policy support, a base on national and international law that is against terrorist organizations, with frameworks to plan, prepare and execute OCO targeting terrorist cyber-infrastructure, is an example of low-entry/low-risk OCO that generates much experience.

A model of maturity development to conduct offensive cyberspace operations for cyber security requires a combination of a deterrence strategy with a spoken component that states the use of OCO for cyber security; a policy where policy-makers feel comfortable and take the necessary political risk; an interpretation of international law which provides much leeway for OCO and Cyber-ISR; frameworks for planning, preparing and executing OCO and Cyber-ISR, but also frameworks that measure the policies. Diplomatic Dialogue is a back-channel to ensure communication between capitals. When states cross red lines, Diplomatic Dialogue enables Heads of State to discuss and lower tensions, just like in the Colonial Pipeline case.

References

Fischerkeller, M. P., & Harknett, R. J. (2017). Deterrence is Not a Credible Strategy for Cyberspace. *Orbis, 61*(3), 381–393. https://doi.org/10.1016/j.orbis.2017.05.003

Healey, J. (2019). The implications of persistent (and permanent) engagement in cyberspace. *Journal of Cybersecurity, 5*(1), 1–15. https://doi.org/10.1093/cybsec/tyz008

Healey, J., & Jenkins, N. (2019). Rough-And-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing. *International Conference on Cyber Conflict, CYCON, 2019-May*, 1–20. https://doi.org/10.23919/CYCON.2019.8756890

Huskaj, G. and Wilson, R. L. (2020). An anticipatory ethical analysis of offensive cyberspace operations. *Proceedings of the 15th International Conference on Cyber Warfare and Security, ICCWS 2020,* 512–520. https://doi.org/10.34190/ICCWS.20.054.

Kosseff, J. (2019). The Contours of "Defend Forward" under International Law. *International Conference on Cyber Conflict, CYCON, 2019-May*, 1–13. https://doi.org/10.23919/CYCON.2019.8757141

Lewis, J. A. (2018). *RETALIATION American perspectives on international cybersecurity*. 252–259.

Sebekin, S. (2020). Choosing between persistent engagement and deterrence in the American cybersecurity strategy. *Mezhdunarodnye Protsessy, 18*(3), 96–125. https://doi.org/10.17994/IT.2020.18.3.62.3

Categorizing Cyber Activity Through an Information-Psychological and Information-technological Perspective, Case Ukraine

Harry Kantola
Finnish National Defence University, Helsinki, Finland

harry.kantola@mil.fi

Abstract: Russian approach to warfighting includes an informational facet. Western hemisphere usually treats cyber activity as a tool similar to traditional warfighting tools such as rifles, artillery and tanks, whereas the Russian approach has an informational and narrative stance to the whole conflict. Placing information in the focus, switches the cyber activity to serve either an informational-psychological or an informational-technical approach. Examining the activity from this non-conventional trait and correlating it to other activities during the execution, the study highlights the coordination of kinetic and non-kinetic actions in an altered manner. In this article I am examining cyber activity through the terms of information-psychological and information-technological approach to form an understanding of Russian or Russian supported activities in cyber space before and during the Ukraine crisis. This will recognize types of cyber activity connected to actions in the physical environment. Actions identified are categorized and placed in a matrix created on psychological and/or technical clout. From this matrix groups of activities are scrutinized in correlation to other activities to expose possible narratives or underlying themes. The study relies on a variant of Grounded theory and is selected to elude from examine technical methods and actions. The observed timeframe is for the first study from 2021, well before the current hot phase, until summer 2022. This article is the first part of a two-stage study, where the first part examines cyber activity through the terms of information-psychological and information-technological approach. The second study places the previous findings in correlation to actions, reactions and mitigation activities to find out how defensive measures were relevant or if the outcome were result of something else than deliberate defensive (cyber-)activities. Throughout the larger study, the underlying hypothesis is that there is a larger coordination of cyber activities than acknowledged related to the ongoing crisis.

Keywords: Information-Psychological Warfare, Information-Technological Warfare, Cyber Warfare, Cyber Operation Compilation, Cyber Operations

1. Introduction

1.1 Motivation

Technological phenomena and activity within cyber space related to the ongoing Ukraine crisis has been studied in large extent, especially by technical cyber security companies and think-tanks. Common for most of them, is that they focus on observing *how* the activity was carried out technically. A few have explored possible correlations between kinetic activity and cyber events. An example of this is the special report from Microsoft, *Defending Ukraine: Early Lessons from the Cyber war* (Microsoft, June 2022), which exemplifies some activity, where a direct link between different attacks in a physical and artificial environment can be demonstrated.

In addition, purposes and coordination of activities has not been academically scrutinized. Present reports and studies make assumptions either that there is no real cyber activity correlated to the activities in the physical world, since there have not been any large-scale breakdowns in Ukrainian cyber space (Lewis, 2022) or that the sophistication of the attacks is not high enough. Therefore, the assumption is that cyber has not played a large role in the crisis.

Through this study, I aim to prove the above stated assumption incomplete and to present, that there has been an attempt to coordinate activities in cyber space with activity in the physical world. Moreover, I will highlight how different types of activity has been utilized during the observed timeframe. The fact, that the known cyberattacks has not had a major breakthrough has more to do with the overall poor conduct of the atrocities overall than with due to lack of attempt within cyber space.

1.2 Purpose, scope and structure

This research examines cyber activity from an informational context, and not from a technical categorization, as analysis of cyber incidents are usually conducted. Main focus is why observed activities have been performed, not how it technically was conducted.

The research area has been limited to observe activities that are related to the Ukraine crisis and has further more been narrowed down in this paper to include activities just before the main assault took place in February 2022 until late summer 2022. This includes activities that relate to the crisis also in late 2021. For the sake of clarity, a few examples from 2015-2021 have been included in the material.

The structure of the article is as follows. First, I will justify the use of the chosen theory and tools. By this, I will introduce an alternative way of looking at activities in cyber space compared to traditional western approach. This is also to present an introduction to the Russian way of thinking regarding the utilizing cyber activities as part of warfare. Finally, the analysis of the material is considered and conclusions are drawn from the processed material.

2. Used theory

The ambition with this study is to examine the how cyber activity has played a role in the atrocities related to the ongoing war in Ukraine, explicitly observed by a non-western approach dividing the phenomenon in informational-psychological and information-technical category.

To elude from a purely technical approach I have chosen to use Grounded Theory as a tool to categorize the events in an appropriate way for the study. The purpose is not to examine the technical actions, but to strive for the purpose and connection to other ongoing activities during the aggression. Therefore, it is important to fade the normal technical approach and catalogue the actions differently.

The purpose of Grounded Theory is to develop theoretical explanations about a specified phenomenon. (Creswell, 1998). Through this inductive analysis the product, patterns, themes and categories, would emerge out of the data rather than being imposed on them prior to data collection and analysis. The focus is to gather data from a variety of sources. As the data is scrutinized by the researcher, themes are expected to emerge (Bowen, 2006). A central principle of data analysis is a constant comparison of the data with other similar groups of data to find similarities and differences. Through this reoccurring comparison trends and patterns are being refined. The used method described is deviated from Strauss and Corbin original axial coding (Bryman 2008).

2.1 Grounded Theory as a tool

As stated earlier, the aim for this study is to find patterns and correlations on actions taken in cyber space and other activity specially to find larger groups of activities that can then be compared to Russian military doctrine and concepts. Emerging groups, patterns and categories are compared to each other and finally compared how these groups and categories correlate to the categorization of information-psychical and information-technological thinking.

Accordingly, to the theory, coding the data, regrouping and then scrutinizing the data produces concepts and categories. The following outcome would be a hypothesis regarding the observed phenomenon, when the data is further processed. (Bryman, 2008). The found concepts and categories have been of importance to be able to compare, correlate and observe relations to actions in the physical domains.

2.2 Suitability of grounded theory in observing cyber activity

Regrouping conducted activities obscures the technical details and makes it easier to find out the method of attack. For the sake of this study, it is not necessary to study these methods in the detail, but rather when these are used during military operation phases. Thus, the important thing is to find out types of action, when those are used and how it correlates to doctrine and actions in other domains.

Even though Grounded theory is originally a sociological approach, it has proven to be suitable to extract the needed information for this analysis. This study has found indicators of used concept and existing categories and created a preliminary hypothesis regarding the ways and how activities have been conducted in coordination with regular warfare in the Ukrainian crisis during 2022.

3. Alternative approach to coordinate cyber activity for state purposes

Russian information security doctrine (Доктрина информационной безопасности Российской Федерации [Doktrina], 2016) and lower regulations and guidelines in accordance with it determines how Russia handles cyber space. A notably difference in terminology is that in Russian language there is no “cyber space”. Instead, the terminology uses “information space” as the closest equivalent (Ristolainen 2017).

Keir Giles states that there is a difference in how cyber capabilities are treated in western and eastern approaches to information and cyber activities. Whereas western society threats cyber activity as a separate function, or nowadays as a domain, Russian consider these activities as just tools in the larger information ecosystem. (Giles, 2016). This consideration applies also for Chinese categorization. (Thomas, 2004)

The phenomena which are considered to take place in cyber space, is referred to as "information space" in Russian terminology. These activities are further divided into information-technical and information-psychological strands in Russian thinking. (Giles, 2016). "*Warfare in information space can be information-technical, when informational technical systems are objects of influence in cyber space, or it can be information-psychological, when the adversary tries to influence a person's mind, his or her moral and mental world, political opinions and ability to make decisions*" (Kari, 2019)

Russia had an attempt to introduce "electronic Russia" already in 2001, (Thomas 2010), but the lessons learnt in the Russian-Georgian conflict 2008 reshaped the approach to the traditional approach. The conflict pinpointed the need to better control the informational aspect of war and coordinate the activities with the traditional domains. This includes the control of information-psychological and information-technological undertakings in external, internal and military activities. (Thomas, 2010)

Russian military doctrine practically dates back to 2010 (Security Council of the Russian Federation), although minor changes were made to it in 2014. In accordance to these doctrines and follow up guidelines, Russia has to be able to inflict in both cognitive and physical space. Combining the doctrine and guidelines for the military with the existing information doctrine highlights, that the Russian approach is divided in two; information-psychological and information-technological approach.

It has been taken into account, that the Russians do not have actual cyber-doctrine. Instead, there is an information doctrine that includes also issues, which "in western terminology" would be described as actions in cyber space. Examining the activities in Ukraine through the alternative approach might give a better understanding on how and why activities has been Implemented.

4. Case Ukraine

4.1 Dataset

Cyber activity has not been widely discussed during the conflict. This might partly be related to the fact that there have not been any large-scale knock-on effects on human lives whereas the conventional activity has. (Lewis, 2022). Additionally, there were a critical uncovering of malicious cyber activity just in the last minute before the attack, when Mandiant discovered the wiping malware aimed at Ukraine. (Mandiant, 2022)

Up until summer 2022 there has been reported far beyond 300 separated cyber incidents in the Ukrainian crisis (Microsoft, June 2022). In this study I have limited the examination to focus on the year 2022, including activities in January and February before the conventional attack was initiated. The material has been extended to with some relevant additions from the timeframe starting with 2015 to 2021 due to their direct linkage to the actions in 2022. The dataset has been amassed from different databases, such as national Cert databases, and cybersecurity companies' factsheets or bulletins. These have been supplemented with relevant news releases (see references).

Data from far beyond 350 separate reports regarding cyberattacks has been filed per technical indicators. These has then been regrouped and merged to form attack type categories, especially to remove duplicate reports, where the same event is reported under a different name. A typical example of this is the HermeticWiper (SentinelLabs), aka. FOXBLADE (Microsoft), aka. KillDisk (Mcafee), malware, that has different name depending on which cybersecurity company is reporting it. The coding process produced a set of 90 categories, which have then been subordinated to constant comparison and further more correlation examination to both actions on the ground and relation to existing military doctrine.

The number of attacks carried out indicates that the use of cyber has been incorporated as a tool in the toolbox. Missions have been conducted both to inflict on critical infrastructure and command and control functions as well as affecting accessibility to information and media outlets.

4.2 Examination of the data

The material shows that there Is a specific approach where some types of attacks relate to information-technical actions connected to events in the physical world and some types of attacks are meant to affect the cognitive aspects of the victim or victims. These strikes wary pending on type of target and the desired effect.

In order to examine how the attacks have been carried out in different categories, I have divided the material into a grid that is divided first of all into information-technical and information-psychological divisions and

secondly in an assessment of their severity. Picture 1 will depict a snapshot of how the material has been categorized and regrouped accordingly to the used grounded theory. The picture shows activities early

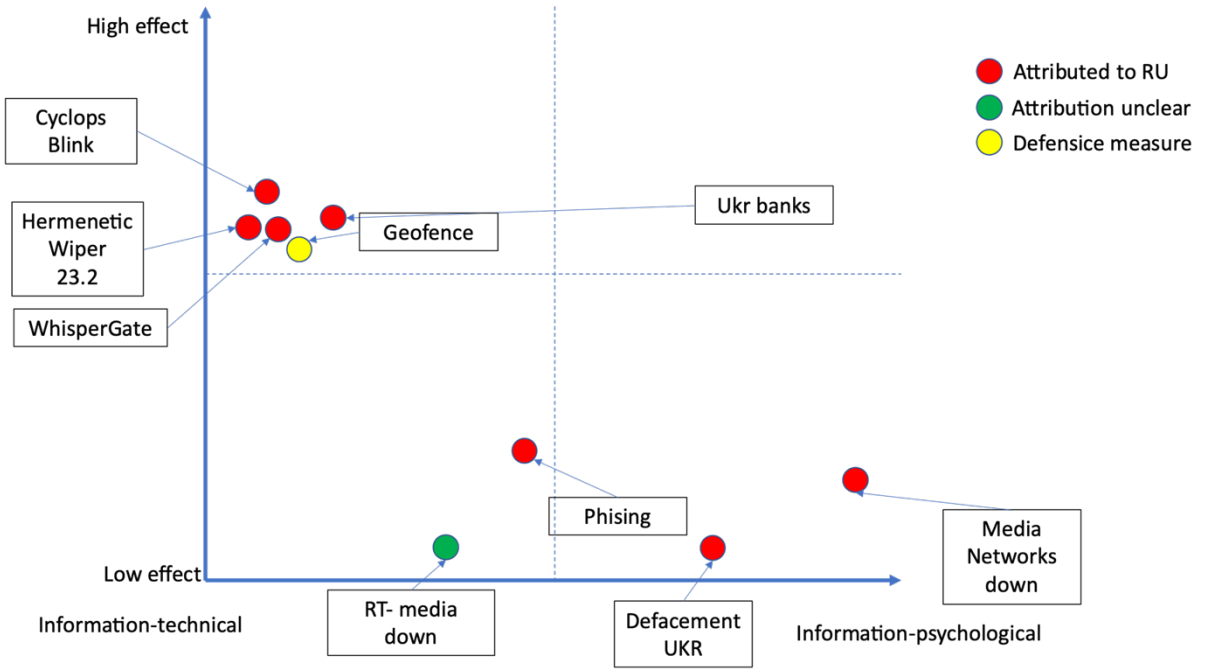


Figure 1: Snapshot of the categorization work

This material has after categorization then been examined from different perspectives. The only common unitary factor, is the division into information-psychological and information-technological classes in the processing of the materials.

The scale and criteria used in this study is entirely created for this purpose. The upper limit of the scale us set to the level 5. This level describes a potentiality for total destruction of the system or impact on the decision making or society. On the other end, nuisance is given the level 1. The scale in between represents different levels of potential impact. An impact level 2 represents effects, that have a short time impact or effects that can be repaired and corrected within a reasonable time. Typical and normal resilience measures should ward of these types of attacks. Impact level 3 represents attacks that potentially requires defensive actions or counter measures. Severity of the potential impact rises above minor damage either in the system or within the situational awareness or decision cycle. Finally, an impact level 4 describes attacks that potentially can severely harm actions, operations, decision making and/or the society in at a length or requires substantial efforts to tackle. All examined attacks are given a level based on their potential effect or outcome, regardless if the attack were successful or not.

Examining the events by severity, it can be noted that information-technological attacks have often been able to create more serious ramifications. Although not all attacks achieve a high penetration into systems, these have had an impact on the comprehensive situation in the area. As a rule, information-technical cases achieve a higher level of effectiveness, although not all of them exceed the level of "teasing".

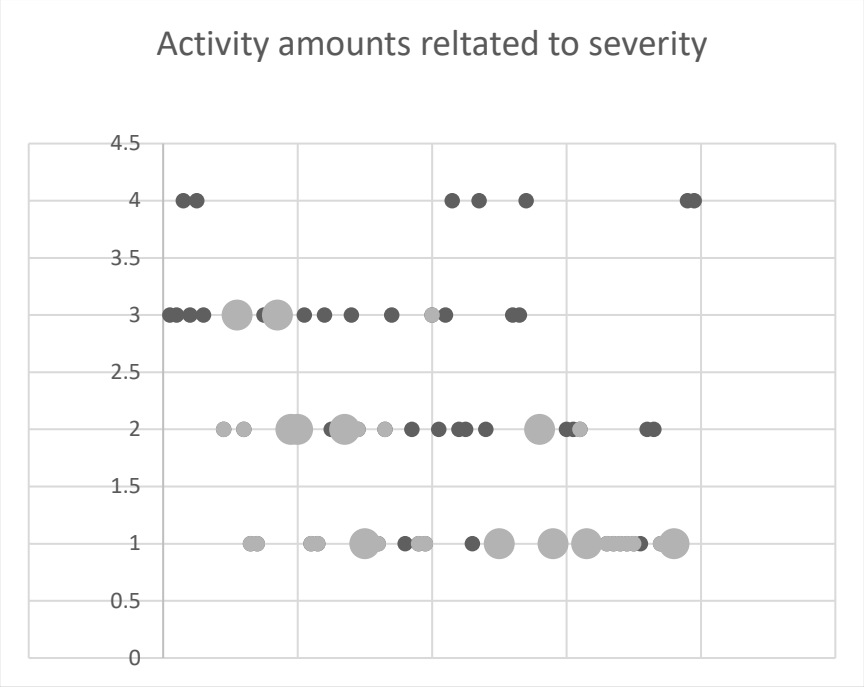


Figure 2: Categorizing the material according to severity. Dark dots represent information-technical activity and light grey dots information-psychological. Size of dot represent amounts of activity.

At the same time, I note that information-psychological events tend to be carried out more frequently. Information-psychological attacks are also larger in number and also targeting more actors at the same time. Thus, as the scope of targets are easily increased, the severity does not increase nor have an impact on the target audience as a whole.

I also note, that the information-technological events took place earlier in the crisis and the information-psychological actions relatively evenly throughout the observed timeframe. This can be identified more clearly in picture 3, where the line presents the relation between the two observed categories.

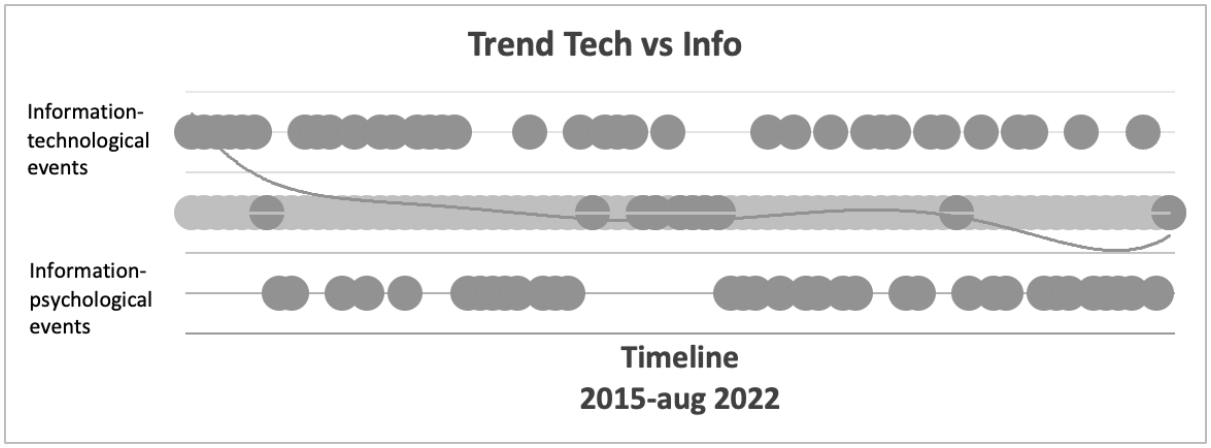


Figure 3: Information-technical and information-psychological activities on a timeline.

The preliminary assessment also suggests that as while the technological level shifts towards less sophisticated targets, it also seems to change to a greater extent opportunistic approach. This seems to be true for both information-technological and information-psychological objectives.

5. Conclusions, discussions and further work

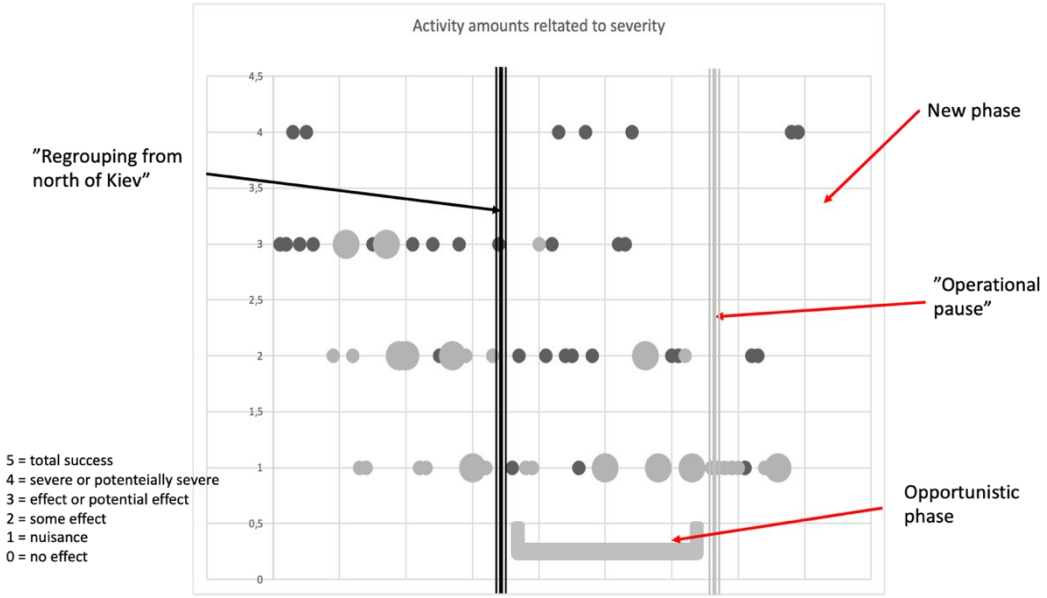
5.1 Conclusions

The result of the study shows first of all, that there was originally a plan to combine kinetic and non-kinetic action. The non-kinetic actions were aimed at both information-technical systems, that is the function of the system, and psychological, that is the cognitive level of citizens and governmental actors through cyber space. Higher in severity and more technical approach were conducted early in the conflict while actions towards the cognitive domain were more of a supporting role. Attacks conducted in the information-psychological sphere were to add on to the fog of war and ability to have a clear picture of what is going on. In addition, the information-psychological approach aimed apparently to change the mindset of the citizens, after a victorious short kinetic battle.

Secondly, it can also be seen from the material that while the actions in the physical world did not work out as planned, the correlating planned actions towards technological systems were (or in fact, tried to be) carried out in preplanned manner. The fact that a large part of the activity was repelled, does not have to do with the events in the physical world, but more with that many attack vectors had been tried before or the malware were found in due time by defenders.

Thirdly, when the planned activity in the traditional domains and cyber space did not work, the activity shifted from information-technical towards information-psychological activity. At the same time the attacks seem to be more and more carried out opportunistically. The data shows also that within the information-psychological niche there were attacks conducted accordingly to premade plan and ad-hoc new attempts performed when the progress with the boots on the ground did not proceed as planes.

Furthermore, the activity in cyber space coincides with the activity on the ground. When the Russians had to rearrange (withdraw) from the northern attack direction and to establish what is called in military terminology an operational pause, there were also an “operational pause” in the activity in cyber space. The above-mentioned analysis and information are depicted in picture 4.



Picture 4: Operational phases of land battle correlated with cyber activity.

After the initial attack and the first phase of the war the connection and coordination regarding attacks in the physical world and in the cyber domain decline. There are still attempts to create more value combining actions, but these are only sporadic incidents. Also new information-technological attacks have been conducted after an extensive preparation phase. These are less in amount and less in coordination with other actions taken. Common for both approaches, is that the severity has gone down over time. Only some sporadic attempts have had a severe potentiality, especially after an extensive preparation phase at the very end of the observed timeframe.

5.2 Discussion

Examining the material gives a suggestion that the planned and effective actions were not conducted in planned time. Possibly due to changes in plans and timing. This led to exposure and disclosure. A large amount of the more severe information-technical attacks could be ward off in the nick of time, due to the actions the dormant malwares conducted while waiting for the right time. The preliminary analysis suggest that the malwares were too early in place and did not get permission to launch as planned.

The material indicates also that there has been a strategic masterplan to combine the action in the kinetic and non-kinetic world. The fact that it didn't succeed has more to do with the struggle in real life and that the war has not gone after plan, than with poor coordination between the domains. This shows that coordinating activities in real time challenging.

Examined cases demonstrates that Russian approach has a clear division in information-technical and information-psychological activities. The Russian forces tries to utilize both approaches. Each of the methods has a role within the military actions and has to be included in the military strategies and operations. As stated before, the challenge is to be in due time.

Actions in cyber space, preparation, tests and footholds, can give indications on upcoming activities in physical world. As seen in this case study, activities in late 2021 and January 2022 gave indications away. The same giveaway came when malwares started to beacon for permission to act. All these gave the defenders tools to repel forthcoming attacks.

At this stage, it is too early to draw final conclusions. It seems to be indications that traditional doctrine is implemented in cyber- (information) actions and goal settings. Activities had been planned in conjunctions with assessed advance on the ground.

5.3 Further work

Elaboration, regarding how much effort were placed on changing citizens mindset versus affecting havoc on systems needs further examination. Fast and deliberate initial attack on systems and a shift towards cognitive influence has still to be proven, even though the material indicates a correlation.

By expanding data set to include pre 2014 activities, there should be possibilities to find out how much of indications and warnings, mitigation techniques and defense measures were put in place before the actual attack in February. This analysis would also bring to light how much of same techniques, malwares, and vulnerabilities were utilized or reused. For this purpose, the timespan should be expanded to include data from the years 2013 to 2022.

Found data indicates also that by analyzing selected timeframes, there might be possibilities to find reasons why imminent attacks were revealed. However, this information requires further investigation and study.

References

Avertium (2022), *CISA Warns of renewed Russina Threat as new Activity is seen in Ukraine,* Avertium, https://www.avertium.com/blog/cisa-warns-of-renewed-russian-threat-new-activity-seen-in-ukraine

Black, D. (March, 2022), *Hero hackers claim to have breached Belarusian weapons firm,* Cybernews, https://cybernews.com/news/hero-hackers-claim-to-have-breached-belarusian-weapons-firm/

Black, D. (April 2022) *Ukraine left reeling by 'zombie' cyberattacks*, Cybernew,s https://cybernews.com/cyber-war/ukraine-left-reeling-by-zombie-cyberattacks/

Bowen, G. A. (2006) *Grounded Theory and Sensitizing Concapts,* Ingernational Journal of Qualitative Methods.

Bryman A. (2008) *Social research methods (third edition)*. Oxford University Press, New York, ISBN 978-0-19-920295-9

Cert-UA (2022), *Новини, (Eng. news), Ukraine state, https://cert.gov.ua/articles*

Creswell, J. W., (1998) *Qualitative Inquiry and Research Designe: Chooose Among Five Traditions.* Sage Publications. London

Corfeld, G. (2022), *Ukraine shrugs off mass govt website defacement as world turns to stare at Russia,* https://www.theregister.com/2022/01/14/ukraine_cyberattack_gov_websites_defaced/

Doktrina (2016) *Doktrina infomatsionnoi bezopasnosti Rossiiskoi Federatsii, [Доктрина информационной безопасности Российской Федерации]*, (information security doctrine of the Russian Federation), http://kremlin.ru/acts/bank/41460/page/1 viewed 9 august 2022,

ESET Research, (April, 2022) *Industroyer2: Industroyer reloaded,* WeLiveSecurity, https://www.welivesecurity.com/2022/04/12/industroyer2-industroyer-reloaded/

ESET Research (May, 2022), *Sandworm uses a new version of ArguePatch to attack targets in Ukraine, WeLiveSecurity,* https://www.welivesecurity.com/2022/05/20/sandworm-ukraine-new-version-arguepatch-malware-loader/

Fisher, D., (2022), *Sandworm group deploying new Cyclops Blinck malware*, Duo Security – Dechiper, https://duo.com/decipher/sandworm-group-deploying-new-cyclops-blink-malware
Gatlan, S. (September, 2022) *Google says former Conti ransomware members now attack Ukraine*, Bleeping computer, https://www.bleepingcomputer.com/news/security/russian-sberbank-says-it-s-facing-massive-waves-of-ddos-attacks/amp/
Geller, E. (2022), *Ukraine prepares to remove data from Russia's reach,* Politico https://www.politico.com/news/2022/02/22/ukraine-centralized-its-data-after-the-last-russian-invasion-now-it-may-need-to-evacuate-it-00010777
Giles, K. (2016) *Handbook of Russian Information Warfare,* NATO Defence College Fellowship Monograph Series 9 ISBN978-88-96898-16-1
Github (2022). Github.com https://github.com/Orange-Cyberdefense/russia-ukraine_IOCs/pulse
Greenberg, A. (2022) *Destructive Hacks Against Ukraine Echo Its Last Cyberwar,* https://www.wired.com/story/russia-ukraine-destructive-cyberattacks-ransomware-data-wiper/
Hacker News, (2022), *Russian Hackers APT28 and UAC-0098 Target Ukraine Again,* https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/
Haskell-Dowland, P. (2022), *As Russia wages cyber war against Ukraine, here's how Australia (and the rest of the world) could suffer collateral damage*, The Conversation; academic rigour, journalistic flair, https://www.helpnetsecurity.com/2022/02/24/cyber-attacks-ukraine/
Holland, S., Pearson, J:, (2022) *US, UK: Russia responsible for cyberattack against Ukrainian banks*, https://www.reuters.com/world/us-says-russia-was-responsible-cyberattack-against-ukrainian-banks-2022-02-18/
Jones, D. (2022). *Botnets, data wiping malware spread as Ukraine incursion begins*, https://www.cybersecuritydive.com/news/botnets-data-wiping-malware-ukraine/619362/
Jonsson, O. (2019) *Russian information warfare and its challenges to international law,* Routledge Handbook of war.
Kari, M. J., (2019), *Russian Strategic Culturi in Cyberspace; Theory of Strategic Culture; a tool to explain Russia's Cyber Threat Perception and Response to Cyber Threats,* Dissertations 122, Jyväskylä University
Kundalyja, D. (2022) Ukrainian hacker leaks Conti ransomware internal chats after gang sides with Russia, Computing, https://www.computing.co.uk/news/4045605/ukrainian-hacker-leaks-conti-ransomware-internal-chats-gang-russia
Leonard B. (2022), *Continued cyber activity in Eastern Europe observed by TAG*, https://blog.google/threat-analysis-group/continued-cyber-activity-in-eastern-europe-observed-by-tag/
Libicki, M. (2007), *Conquests in Cyberspace, National Security and Information Warfare*, Rand Corporation, Cambridge University Press. ISB*N* 978-0-521-87160-0
Lewis, J. A. (2022), *Cyber War and Ukraine*, Centre for Strategic and International Studies, report on: https://www.csis.org/analysis/cyber-war-and-ukraine
Mackie, K. (2022) Microsoft Ukraine Report Warns of Coming Zero-Day Exploits, Redmond, https://redmondmag.com/articles/2022/04/27/microsoft-ukraine-report-warns-of-coming-zero-day-exploits.aspx
Malwarebytes lab, (2022), *Cobalt Strikes again: UAC-0056 continues to target Ukraine in its latest campaign,* https://www.malwarebytes.com/blog/threat-intelligence/2022/07/cobalt-strikes-again-uac-0056-continues-to-target-ukraine-in-its-latest-campaign
Microsoft Corporation (April 2022), *Special Report Ukraine: An overview of Russia's cyberattack activity in Ukraine.* https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Vwwd
Microsoft Corporation (June 2022), Defending Ukraine: Early Lessons from the Cyber War, *https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE50KOK*
Milmo D. (2022), *Anonymous: the hacker collective that has declared cyberwar on Russia*, The Guardian, https://www.theguardian.com/world/2022/feb/27/anonymous-the-hacker-collective-that-has-declared-cyberwar-on-russia
National Cyber Security Center (UK), (2022), https://www.ncsc.gov.uk/news/uk-organisations-encouraged-to-take-action-around-ukraine-situation
Petkauskas V. (2022) *Eurovision cyberattack: pro-Russian hackers declared 'war' on ten states,* https://cybernews.com/cyber-war/eurovision-cyberattack-pro-russian-hackers-declared-war-on-ten-states/
Ristolainen, M. (2017): *Should 'RuNet 2020' be Taken Seriously? Contradictory Views about Cybersecurity between Russia and the West*, Journal on Information Warfare, vol. 16, no. 4, 113-131
Schneier, B. (2022) Microsoft Issues Report of Russian Cyberattacks against Ukraine, https://www.schneier.com/blog/archives/2022/04/microsoft-issues-report-of-russian-cyberattacks-against-ukraine.html
Sharma, Y. (2022) *Anonymous leaks 15,600 new emails from GUOV I GS via DDoSecrets*, https://www.thetechoutlook.com/news/technology/anonymous-leaks-15600-new-emails-from-guov-i-gs-via-ddosecrets/
Smith, B. (2022), *Defending Ukraine: Early Lessons from the Cyber War,* Microsoft, https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/
Socradar.io (2022) *What you need to know about Russian Cyber escalation in Ukraine; The Second Wave of Cyber Attacks on Ukraine: Deadlier than the First Wave.* https://socradar.io/what-you-need-to-know-about-russian-cyber-escalation-in-ukraine/#The-Second-Wave-of-Cyber-Attacks-on-Ukraine-Deadlier-than-the-First-Wave

Symantec (February 2022), *Ukraine: Disk-wiping Attacks Precede Russian Invasion*, https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/ukraine-wiper-malware-russia
Symantec (April 2022), *Shuckworm: Espionage Group Continues Intense Campaign Against Ukraine*, https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/shuckworm-intense-campaign-ukraine
Telychko, V. (2022), AgentTesla Information-Stealing Malware Delivered in Cyber-Attacks on Ukrainian Government Entities, SOCPRIME, https://socprime.com/blog/agenttesla-information-stealing-malware-delivered-in-cyber-attacks-on-ukrainian-government-entities/
Thomas, T. L. (2004) *Russian and Chinese information warfare: theory and Practice*, Presentation, https://apps.dtic.mil/sti/pdfs/ADA467510.pdf)
Thomas T.L (2010) *"Russain Information warfare Theory: The consequences of August 2008" in s Blank and Weits (eds.) The Russian military today and tomorrow,* 2010.
Toulas, B (April, 2022), *Russian hacktivists launch DDoS attacks on Romanian govt sites,* Bleeping computer, https://www.bleepingcomputer.com/news/security/russian-hacktivists-launch-ddos-attacks-on-romanian-govt-sites/amp/
Toulas, B. (May, 2022). *Russian Sberbank says it's facing massive waves of DDoS attacks,* Bleeping computer, https://www.bleepingcomputer.com/news/security/russian-sberbank-says-it-s-facing-massive-waves-of-ddos-attacks/amp/
Toulas, B. (June, 2022) *Russian govt hackers hit Ukraine with Cobalt Strike, CredoMap malware*, Bleeping computer LLC, https://www.bleepingcomputer.com/news/security/russian-govt-hackers-hit-ukraine-with-cobalt-strike-credomap-malware/
Treloar, S. (2022). *Russian Hackers Target Norway in Latest Volley of Cyber Attacks*, Bloomberg, https://www.bloomberg.com/news/articles/2022-06-30/russian-hackers-target-norway-in-latest-volley-of-cyber-attacks
YouTube, (2022), https://www.youtube.com/watch?v=dwbDMnEraO0, (at 5 min 23 sec)
Zetter, K. (2022), *Russia Began Setting Stage for Cyberattacks Against Ukraine a Year Ago,* Zero Day, https://zetter.substack.com/p/russia-began-setting-stage-for-cyberattacks?s=r
Zorz, Z. (2022) *Cyber attacks on Ukraine: DDoS, new data wiper, cloned websites, and Cyclops Blink,* HelpNetSecurity, https://www.helpnetsecurity.com/2022/02/24/cyber-attacks-ukraine/

Using Military Cyber Operations as a Deterrent

Maria Keinonen
National Defence University, Helsinki, Finland

maria.keinonen@mil.fi

Abstract: The Deterrence theory was formed after the World War II to describe the tensions between nuclear-armed states. Because of its origins, deterrence is mainly researched from the point of view of powerful states. However, deterrence nowadays is essential for any state to include in their strategies. The ever-increasing dependence on technology forces states to protect their sovereignty in cyberspace as well as in other domains. Therefore, cyber operations should be considered not just as a means to protect the cyber domain, but as means of deterrence. Cross domain deterrence (CDD) is a theory that includes all the warfighting domains in creating deterrence, including cyberspace. Despite these new perspectives, the use of military cyber operations as a deterrent has been studied mainly in terms of offensive strategies. Therefore, incorporating all types of military cyber operations into deterrence strategies is understudied. This study focuses on the possibilities of a small state to use cyber operations to create deterrence. The research question is: "How can a small state use cyber operations as a deterrent?" According to the Finnish understanding, cyber operations can be divided into three types: offensive, defensive and supportive operations. These types of operations each have their separate role in securing the cyber domain and freedom of action in every military domain, as well as influencing the opponent´s ability to act as planned. Using Finland as a case study, this paper argues that using military cyber operations is noteworthy for any state dependent on cyberspace, not only for military purposes, but for building CDD. The CDD theory and characteristics of cyber operations are studied in order to form better understanding of the topic and provide ideas for academic discussion. The research methods are content and SWOT analysis. The key observation presented is that each type of cyber operation has a role in forming CDD. For a small state, it´s profitable to use every type of cyber operations and thus expand the tool box for deterrence.

Keywords: Cross Domain Deterrence, Cyber Deterrence, Finland, Offensive Cyber Operation, Defensive Cyber Operation, C5 operation, Cyber Resilience

1. Introduction

The core idea of deterrence can be summed up as an effort to convince the other party that a hostile action is not profitable and both can achieve their goals without provoking the other into conflict (Mazarr & Goodby, 2010). For a small state this is a continuous battle of being credible enough, but not too provocative in the eyes of a bigger adversary. It can be questioned, if a small state has the ability to build deterrence with a credible threat (Hanska, 2019) or if deterrence is in general more of a bargaining process with something valuable for both parties to gain or lose (Kerttunen, 2019).

The Deterrence theory was formed after the World War II to describe the tensions between nuclear-armed states. It focused purely on the use of military force, especially nuclear weapons, and its prevention (Morgan, 2003). With the development of technology and the threat environment becoming more diversified with threat actors, the classic way of structuring deterrence was no longer sufficient. (Arie, 2016). The toolbox of deterrence must be expanded including political, economic, diplomatic and military means. (Sweijs, T., Zilincik, 2021)

Because of its origins, deterrence is mainly researched from the point of view of powerful states. However, deterrence nowadays is essential for any state to include in their strategies. The ever-increasing dependence on technology forces states to protect their sovereignty in cyberspace as well as in other domains. This paper focuses on the possibilities of a small state to use cyber operations to create deterrence. The research question is: "How can a small state use cyber operations as a deterrent?"

The concept of a small state can be addressed via materialistic values, as relatively smaller territories, populations and resources than surrounding neighbours (Radoman, 2018). It can also be considered as a deficit in influence and autonomy in position of international relationships (Goetsel, 2000). Small states are relatively more vulnerable for international security threats with limited options to mitigate them compared to their bigger counterparties (Väyrynen, 1997). Small states are expected to favour international cooperation that strengthens common norms and institutions in international security (Coolsaet, 2004).

A small state in this paper refers to a state with relatively lower population and fewer resources than its neighbouring countries and with an intention to strengthen international norms to stabilize the global threat environment. A small state must use every cost-effective means of creating deterrence, even if the end result is uncertain (Toomse, 2015). This research understands cost as a combination of time, money, human recourses and technology spent to achieve a goal.

Finland has been chosen as the subject of the article's case analysis, because it meets the aforementioned criteria of a small country. Finland shares more than a thousand kilometres of border with Russia, and for more than a hundred years it has managed its relations with its more powerful neighbour, without a military alliance. This changed in the spring of 2022 after Finland applied to become a member of NATO, but the country's security policy solutions have nevertheless been shaped during a time when Finland has remained militarily non-aligned throughout its independence.

2. Background

2.1 Cross domain deterrence

Cross domain deterrence (CDD) can be understood as the use of a threat or a combination of them to dissuade a target from taking actions of any type (Lindsay & Gartzke, 2019). CDD can be viewed as the use of military force in all domains of warfare. An attack in one domain can be answered by action in another domain, so warfare does not have to be completely symmetrical. A broader approach includes non-military elements to the CDD theory, such as political means and economic sanctions. (Sweijs & Zilincik, 2021)

This paper refers to military domains (air, land, sea, space, cyber) as “domains” and. Military domains together with non-military elements (diplomatic, information, economic, legal) are referred as “dimensions” according to a DIMEL (Diplomatic, Information, Military, Economic, Legal) model (Sweijs et al, 2021). Military operations, including cyber operations, are considered to be set of actions in military domains to achieve a mission or goal set for a military force (JP 3-0, 2017). Cyberspace is referred to the cyber environment of the whole society including the military cyber domain.

A comprehensive approach might create in smaller states an opportunity to exercise deterrence without risking too much in provoking bigger adversary, especially while acting with other states with similar goals (Toomse, 2015). An example of this is Finland participating in European Union economic sanctions against Russia because of the war in Ukraine.

The CDD theory broadens the use of available capabilities to make symmetric or asymmetric moves. It increases the synergy between these capabilities and creates a portfolio for deterrence options. It also commits society's actors more broadly to building deterrence. (Lindsay & Gartzke, 2019) The ability to create asymmetry might act as a benefit for a small state, since it´s not always profitable to face a greater power with similar means. Also, involving actors in the whole society adds resources to building deterrence.

It could be asked if CDD diminishes the importance of deterrence by broadening the view beyond the use of nuclear and military power and is thus doomed to failure, because the softer means are less likely to deter threats. (Gartzke & Lindsay, 2017a) On the other hand, deterrence is no longer see as an absolute concept as it was in the last century. Rather, it is a competition where states take turns climbing the escalation steps with their opponents. (Lindsay & Gartzke, 2017b)

CDD's idea that a threat in one domain can also be countered in another domain is also well suited to deter threats in cyberspace. Combating cyber threats with cyber means alone can never be complete, since building cyber defence is slow and expensive compared to executing cyberattacks. Inevitably, there will always be attack vectors to be exploited in the cyberspace. (Taddeo, 2018a) Instead of countering every cyberattack possible, it is more efficient to draw the line between what is absolutely reprehensible and will cause countermeasures (Rivera, 2015). This could increase the credibility of deterrence.

2.2 The role of military cyber operations in Finland

Finland divides cyber operations into three types: offensive cyber operations (OCO), defensive cyber operations (DCO) and Command, Control, Communications, Computers, Collaboration operations (C5O). Cyber operations also involve control of the battle space, intelligence and modification of the battle space, just like any other military operation, but instead of the physical domains, the actions are aimed at the cyber domain. These are actions supporting cyber operations. (Laari et al, 2019)

Defensive cyber operations protect friendly cyber domain and they are carried out against a specific threat. The goal is to prevent or disrupt the adversary's actions and restore the safety of compromised cyberspace. Protective measures (Defensive Cyber Operation Internal Defence Measures, DCO-IDM) consist of active and proactive means to detect and respond to a threat factor within the friendly cyberspace. The countermeasures

(Defensive Cyber Operation Responsive Actions, DCO-RA) extend beyond the cyberspace to be protected. It requires that the origin of the attack has been verified. (Laari et al, 2019)

The goal of offensive cyber operations is to project power in or through a foreign cyberspace to support one's own goals. OCO can target targets in the adversary's cyberspace or cause effects in the physical domain. (Laari et al, 2019)

C5 operations secure and maintain elements of friendly cyber domain. Unlike defensive cyber operations, C5O are not carried out against any specific threat, but are used to prepare against all threats that weaken friendly cyber domain. C5O can be understood as a continuous operation for cyber security maintenance. (Laari et al, 2019)

In Finland, the statutory duties of the Finnish Defence Forces (FDF) require the defence of national security in the event of a military attack (11.5.2007/551, 2007), and this also applies to the defence of state sovereignty in the cyberspace against state actors (Finnish Government, 2021). Military cyber defence capability consists of intelligence, offensive and defensive capabilities (Puolustusministeriö, 2019).

According to the Governments Defence Report 2021, not only the FDF´s own systems but also other systems that directly affect the national defence capability can be better secured with the military cyber capabilities (Finnish Government, 2021). However, these systems are not defined in more detail in the report. It´s also worth noticing that the FDF protects mainly its own military systems. If a cyber attack is targeted to civilian infrastructure, it takes time to identify and prove the attacker as other state. Therefore, the importance of resilience in society as well as cooperation and information sharing with other authorities cannot be overlooked.

According to the Finnish Cyber Strategy, OCO can be used as a tool for political and economic coercion, and in a serious crisis as one means of influence alongside other traditional military means of force. (Turvallisuuskomitea, 2013) In this case, the actor is a military organization, which must be considered in terms of the potential risk of escalation.

The cyber domain is protected by increasing the threshold for different types of cyberattacks, for example by improving the detection and attribution ability of cyberattacks and the ability for countermeasures. Countermeasures can consist of, for example, law enforcement measures, diplomatic measures or active cyber countermeasures. (Turvallisuuskomitea, 2019)

3. Methodology and results

This study applies characteristics of cyber operations to the theory of cross domain deterrence. Content analysis (Puusa, 2021) was used as a research method so that factors could be separated from the investigated material in order to answer the research question.

The material concerning CDD and cyber deterrence was gathered from the abstract and citation database Scopus. The search was limited to the last ten years, since the perception of cyber phenomena have changed rapidly during past decades. After reaching saturation, some material was excluded due to being less relevant. The material concerning Finland was searched from the websites of Finnish Government and the FDF. The analyzed material contained five Finnish strategic documents, one handbook of Finnish military cyber operations, one US doctrine of cyber operations, five CDD articles and seventeen cyber deterrence articles.

During the first round of analysing, themes supporting the research question were defined as belligerence and cost-effectiveness of cyber operations and deterrence. Finnish cyber operations were investigated to create factors describing the types, functions and actors of cyber operations.

After the initial coding, in order to better understand the characteristics of military cyber operations, a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis (Pöyhönen, 2018) was carried out analysing cyber deterrence literature. The analysis focused on the internal strengths and weaknesses of military cyber operations used to secure Finland's sovereignty, as well as external opportunities and threats. The results are presented in the Table 1.

Based on the cyber deterrence literature, OCO can be executed cheaper and faster than other cyber operations. Since the attack can be planned beforehand at specific targets, it doesn´t require vast resources to execute (Rivera, 2015; Chen, 2017; Chen, 2018; Taddeo, 2018a) and covert operations are possible because of the attribution problem (Chen, 2017; Fischerkeller, 2017). This could create an opportunity with a surprise element and taking initiative.

OCO and DCO-RA are more belligerent in nature than the other cyber operation types, since the target is an opponent system (JP 3-12, 2018; Laari et al, 2019). Therefore, there is a risk of escalation (Schneider 2019). For managing escalation, in theory, scaling the attack intensity is possible (Leuprecht et al, 2019; Burton, 2018) and in an ideal situation both OCO and DCO-RA are repeatable (Mallick, 2021; Harknett & Smeets, 2022) and even reversible (Fischerkeller, 2017).

The advantage of active use of offensive cyber capabilities can also be seen as a weakness if too much of a state's cyber abilities are revealed and a foothold for the opponent system is lost (Schneider, 2019). There might also appear problems with attribution, speed, accuracy and repeatability (Mallick, 2021).

DCO-IDM include active measures for protecting the friendly cyberspace by tracing the origin of the attack and mitigating its effects on friendly networks. Along with C5O, these actions create the basis for deterrence by denial. If DCO-IDM is implemented with artificial intelligence and machine learning, a real-time ability to detect, and identify the attacker can be achieved. (Rivera, 2015) It´s worth noting, however, that in Finland military operations protect mainly the FDF`s cyber domain leaving out the rest of the society. It is possible for other authorities to ask for assistance for specific situations, but the process might take time.

C5O creates the basis for deterrence by denial in creating resilience, because the ability to recover mitigates the effects of cyberattacks. The focus is on the maintenance, management, disruption tolerance, system recovery and development of information systems and networks. The C5O should be understood as a continuous and long-term investment in the construction and development of cyber security. (Laari et al, 2019) In Finland, the FDF protects cyber domain and the resilience of the rest of society is divided with several public and private actors. There´s a risk of deficiencies in information sharing between military actors and other authorities.

Resilience is an important factor of deterrence by denial, but in cyberspace it also requires active means for isolating the attacker and mitigating its effects (Fischer, 2019). This includes an element of deception, which can be achieved by various technical means, such as honeypots or traps of malware that are launched when information is stolen (Wanic & Rowe, 2018). Together with DCO and C5O these passive and active effects can be achieved.

Table 1: SWOT analysis of Finnish cyber operations

Type	Purpose and target	Benefits	Disadvantages	Opportunities	Risks
OCO	Attacking aggressor state outside friendly cyberspace	Credibility, cost-effectiveness, speed	Loosing foothold, revealing capabilities	Gaining surprise and initiative	Escalation, uncontrolled effects, unrepeatability
DCO-RA	Retaliating aggressor actor outside the FDF cyber domain and assisting other authorities when requested	Credibility, cost-effectiveness, speed in FDF networks	Loosing foothold, revealing capabilities, military cyber operations concentrate on the FDF and could be slow on other friendly networks	Mitigating attacks, limiting and isolating of the impact	Escalation, unrepeatability
DCO-IDM	Protecting the FDF cyber domain and assisting other authorities when requested	Activity and anticipation	Inefficiency to focus on essential threats, military cyber operations concentrate on the FDF	Mitigating attacks, deceiving adversary	Lack in performance
C5	Protecting the FDF cyber domain and assisting other authorities when requested	Recovery	Insufficient alone, not feasible by military cyber operations only	Decreasing attack vectors	Cost-effectiveness, passivity, deficiencies in information sharing

After the SWOT analysis, the understanding was deepened with factors found in the CDD theory. The focus on the third round of coding was to define escalation risk and cost-effectiveness of cyber operations and preconditions to be met in order to achieve the desired effects as deterrents. Based on the analysis, a table was created to describe the nature and role of cyber operations as part of cross domain deterrence. The results are presented in Table 2.

From a deterrence point of view, some preconditions must be met before executing OCO. First, red lines should be clearly stated, so there are less chances of mistaking the consequences for hostile actions from the adversary´s point of view (Rivera, 2015). Second, national law and international agreements must be abided by for the OCO to be justified (Huskaj & Moradian, 2018; Rivera, 2015; Schneider, 2019; Mandel, 2017; Fischer, 2019; Wanic & Rowe 2018). Third, in order to be credible, demonstrations of cyber power must be made (Wanic & Rowe, 2018; Taddeo, 2018b). It should also be noted, that the use of other military capabilities could either amplify or moderate the effects of OCO (Fischerkeller, 2017).

The low cost of OCO is based on fact that the execution is fast and requires only little money or human resources compared to defending against an attack (Janczewski & Caelli, 2016). On the other hand, creating the foothold for the attack could take time and require a lot of preparations in the form of information gathering and building a technical capability.

The DCO-RA differs from OCO in its retaliatory nature where OCO can be used in advance. DCO-RA offers means for deterrence by punishment in a milder form than OCO. Countermeasures are used to communicate to the attacker that the attack has been detected and it is possible to respond to it. (Laari et al, 2019)

DCO-RA requires successful attribution (Fischer, 2019) for the retaliation to be justified in the eyes of international actors. There is a risk of this taking too much time for the counteraction to be efficient. This risk can be countered in theory with automated countermeasures (Wanic & Rowe, 2018), that could be facilitated with artificial intelligence (Rivera, 2015; Taddeo, 2018b).

For the DCO-IDM to be successful, the friendly cyberspace needs to be monitored. For this, the FDF needs the cooperation of other domestic and international actors in terms of threat information sharing. If a friendly network outside the FDF needs assistant, for example in terms of threat hunting, there must be a request for help from other authorities. This might take time to execute.

Cyber resilience in Finnish society is built on several actors, including public and private actors as well as citizen (Turvallisuuskomitea, 2013). Therefore, the role of the FDF is minimal regarding any other than its own cyberspace. The military national defence capability is secured with C5O, but the society needs all the actors to cooperate in securing the critical infrastructure. The resilience needs a continuous strategy and efficient resources, that involves the whole society in building and maintaining it.

Table 2: Finnish military cyber operations as deterrents

Escalation risk ↑ / **Cost** ↓

Type	Objective	Deterrence	Preconditions	Benefits and opportunities	Disadvantages and risks
OCO	Attack	By punishment	Law, preparation of targets, continuous foothold, repeatability, signalling red lines	Credibility, cost-effectiveness, speed, gaining surprise and initiative	Loosing foothold, revealing capabilities, escalation, uncontrolled effects, unrepeatability
DCO-RA	Retaliation	By punishment	Law, attribution, signalling red lines	Credibility, cost-effectiveness, speed, mitigating attacks, limiting and isolating of the impact	Escalation, unrepeatability, loosing foothold of opponent system, revealing capabilities, military cyber operations concentrate on FDF
DCO-IDM	Defence	By denial	Surveillance, cooperation	Activity and anticipation, mitigating attacks, deceiving adversary	Lack in performance, inefficiency to focus on essential threats,

Escalation risk	Type	Objective	Deterrence	Preconditions	Benefits and opportunities	Disadvantages and risks	Cost
						military cyber operations concentrate on FDF	
	C5O	Resilience	By denial	Strategy, resources, continuity	Recovery, decreasing attack vectors	Cost-effectiveness, passivity, insufficient alone, not feasible by military cyber operations only	

In the fourth round of the analysis, the gathered information was compared to "A Framework for Cross-Domain Strategies Against Hybrid Threats" by Sweijs, Zilincik, Bekkers and Meessen. The framework divides escalation options into five categories, which can be used in DIMEL (Diplomatic, Information including cyber, Military, Economic, Legal) model with each dimension. Categories are cooperation, persuasion, protection, coercion (including deterrence) and control, increasing in escalatory nature from the first to the last. (Sweijs et al, 2020) This framework is relevant to this study, because it draws to the theory of CDD extending the perspective beyond deterrence. It also addresses the modern threat environment with hybrid threats. For future research, the analysis performed in this paper can be utilized in future research focusing on the framework in its entirety.

This study investigated deterrence options in every dimension and the rest of the framework was excluded. Finnish cyber operations were included in the framework and examples were generated based on the previous analysis rounds. Table 3 sets out the proposal of how Finnish cyber operations could be used as a part of this framework with other dimensions for creating deterrence.

Table 3: Finnish cyber operations in "A Framework for Cross-Domain Strategies Against Hybrid Threats"

Dimension	General examples (Sweijs et al, 2020)	Example of support from dimension to cyber operations	Example support from cyber operations to dimension	CO type related
Diplomatic	Threatening diplomatic isolation in order to maintain the adversary's current behaviour, possibly with diplomatic allies.	Communicating the will to use offensive cyber capabilities and own resilience via strategic documents and official announcements of the state administration.	Demonstrations of offensive capabilities	OCO, DCO-RA, DCO-IDM, C5O
Information	Threatening retaliation through information warfare, counterintelligence	Information gathering and counterintelligence in physical and information domains in order to predict possible cyber attacks	Information gathering and counterintelligence in cyber domain	OCO, DCO-RA, Cyber ISR (Intelligence, Surveillance, Reconnaissance)
Military	Credibly established and communicated retaliatory capability.	Demonstrations of military capabilities	Demonstrations of offensive capabilities	OCO, DCO-RA
Economic	Threatening the use of sanctions/ supply manipulation/ price increase in order to maintain the target's current behaviour.	Sanctions with diplomatic allies targeting supply chains	Demonstrations of offensive capabilities and protecting own critical infrastructure	OCO, DCO-RA, DCO-IDM, C5O
Legal	Threats of legal sanctions to dissuade the adversary from breaking the rule. Alternatively, this includes threats of prosecution within one's domestic legal jurisdiction if a target does not accept one's demands.	Legal support for executing all types of cyber operations in predetermined conditions. Legal support for domestic and international cooperation and information sharing.	Carry out operations in accordance with the law, gathering forensic evidence to support attribution in case of cyber attacks	OCO, DCO-RA, DCO-IDM, C5O

4. Conclusions and future research

As a conclusion, it can be stated that cyber operations, especially C5 and defensive cyber operations, have a role in creating deterrence against threats occurring in or through the cyberspace. Offensive cyber operations work best as part of an overall deterrence that involves not only the use of armed force but also legal, political and, economic means. In order for an OCO to be worthy of being used as a deterrent, there must be a clearly demonstrable reason why this is done. Probably, in such a situation, the threat actor is already doing such provocative and damaging actions that a cyber operation alone is not enough to be used as a sufficient countermeasure, but it is also necessary to use military deterrent measures. At this stage, there may also be a limit as to whether it is purely a deterrence in question, or the prevention of an already starting crisis and the prevention of escalation.

For a small state, using cyber operations as a deterrent from the escalation and cost-effectiveness point of view is bipartite. On the one hand, OCO and DCO-RA might require less resources to implement, but they are also belligerent and could lead to escalation. On the other hand, DCO-IDM and C5 operations require a vast amount of resources to protect the civilian infrastructure and achieving sufficient resilience takes a lot of time, but these means are not escalatory in nature. For a small state´s perspective, it´s necessary to build resilience and profitable to develop offensive and retaliatory cyber capabilities.

In future research, the analysis formed in this article could be extended to include other small states and different ways of structuring cyber operations. To better understand the use of cyber capabilities as a deterrent, it is necessary to investigate the possibilities of states to build a credible deterrence in today's threat environment.

References

11.5.2007/551, (2007). *Laki Puolustusvoimista*, Finlex. https://www.finlex.fi/fi/laki/ajantasa/2007/20070551.

Arie, K. (2016) Complex Deterrence Theory and the Post-Cold War Security Environment, *NIDS Journal of Defense and Security*, No. 17, December 2016, pp. 21-39.

Burton, J., (2018). *Cyber Deterrence: A Comprehensive Approach?* Nato Cooperative Cyber Defence Centre of Excellence, Estonia 2018.

Chen, J. (2017). Deterrence and its Implementation in Cyber Warfare, in *ICCWS 2017 12th International Conference on Cyber Warfare and Security*, United states.

Chen, J (2018). Does Conventional Deterrence Work in the Cyber Domain?, in *ECCWS 2018 17th European Conference on Cyber Warfare and Security V2*, Norway.

Coolsaet, R. (2004). Small States in World Politics: Explaining Foreign Policy Behavior. *Int Polit* 41, 284–285. https://doi.org/10.1057/palgrave.ip.8800051.

Finnish Government, (2021). *Government's Defence Report*, Publications of the Finnish Government 2021:80, Helsinki 2021, ISSN pdf: 2490-0966.

Fischer, M., (2019). The Concept of Deterrence and its Applicability in the Cyber Domain, in *Connections: The Quarterly Journal 18*, no. 1, 69-92. DOI: 10.11610/Connections.18.1-2.05.

Fischerkeller, M., (2017). Incorporating Offensive Cyber Operations into Conventional Deterrence Strategies, *Survival*, 59:1, 103-134, DOI: 10.1080/00396338.2017.1282679.

Gartzke, E., Lindsay, R., (2017a). Thermonuclear cyberwar, *Journal of Cybersecurity*, Volume 3, Issue 1, March 2017, Pages 37–48. https://doi.org/10.1093/cybsec/tyw017.

Goetschel, L. (2000) Small States and the Common Foreign and Security Policy (CFSP) of the EU: A Comparative Analysis, *NRP42 Project-4042-044205 Final Report*, Berne, May 2000.

Hanska, J. (2019). Pelotetta vai pidäkettä? Deterrenssiteorian käytäntöä pienen valtion näkökulmasta. *Tiede Ja Ase*, 2019(1). https://journal.fi/ta/article/view/88681.

Harknett, R. & Smeets, M., (2022). Cyber campaigns and strategic outcomes, in *Journal of Strategic Studies*, 45:4, 534-567, DOI: 10.1080/01402390.2020.1732354.

Huskaj, G. & Moradian, E., (2018). Cyber Deterrence: An Illustration of Implementation, in *13th International Conference on Cyber Warfare and Security,* 8 – 9 March 2018, Washington, DC, USA.

Janczewski, L. & Caelli, W. (2016). Security of small countries: Summary and model in Janczewski, Lech and William Caelli (eds): *Cyber conflicts and small states*, UK, ISBN 978-1-4724-5219-1.

JP 3-0, (2017). *Joint Operations*, US Joint Chiefs of Staff.

JP 3-12*,* (2018). *Cyberspace Operations,* US Joint Chiefs of Staff.

Kerttunen, M. (2019). Beyond Punishment: Deterrence in the Digital Realm. *Connections*, 18(1/2), 61–68. DOI: 10.11610/Connections.18.1-2.04.

Laari, T., Flyktman, J., Härmä, K., Timonen, J. & Tuovinen, J. (2019). *#kyberpuolustus: Kyberkäsikirja Puolustusvoimien henkilöstölle*. National Defence University, Finland. https://urn.fi/URN:ISBN:978-951-25-3120-2.

Leuprecht, C., Szeman, J. & Skillicorn, B., (2019). The Damoclean sword of offensive cyber: Policy uncertainty and collective insecurity, in *Contemporary Security Policy*, 40:3, 382-407, DOI: 10.1080/13523260.2019.1590960.

Lindsay J., Gartzke E. (2017b). Cross-Domain Deterrence and Cybersecurity: The Consequences of Complexity, with Erik Gartzke, in *US National Cybersecurity: International Politics, Concepts and Organization*, ed. Damien van Puyvelde, Aaron F. Brantley (New York: Routledge, 2017). DOI: 10.4324/9781315225623.

Lindsay J., Gartzke E. (2019). Introduction: Cross-Domain Deterrence, From Practice to Theory, in Lindsay J., Gartzke E. (eds) *Cross-Domain Deterrence: Strategy in an Era of Complexity*, Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780190908645.003.0001.

Mallick, P. (2021). *Deterrence Theory– Is it Applicable in Cyber Domain?* The Centre for Land Warfare Studies, Vivekananda International Foundation, New Delhi, ISBN: 978-93-91498-05-4.

Mandel, R. (2017). *Optimizing cyberdeterrence: a comprehensive strategy for preventing foreign cyberattacks.* Washington, DC : Georgetown University Press. ISBN: 9781626164123; 9781626164130.

Mazarr, M., Goodby, J. (2011). Redefining the Role of Deterrence, in Hoover Institution on War, Shultz, R., Drell, G. & Goodby, J. *Deterrence: Its past and future: Papers presented at Hoover Institution,* November 2010. Stanford, Calif.: Hoover Institution Press.

Morgan, P. (2003) *Deterrence Now*. Cambridge: Cambridge University Press (Cambridge Studies in International Relations). DOI: 10.1017/CBO9780511491573.

Puolustusministeriö, (2019). *Kyberpuolustuksen kehittämisen strategiset linjaukset*. ISBN: 978-951-663-069-7 pdf.

Puusa, A. & Juuti, P., (2020). *Laadullisen tutkimuksen näkökulmat ja menetelmät*. Helsinki: Gaudeamus.

Pöyhönen, J. (2018). *SWOT-analyysin soveltaminen yrityksen kyberturvallisuuden tilannekuvan muodostamiseen*. Informaatioteknologian tiedekunnan julkaisuja, No. 58/2018, Jyväskylän yliopisto.

Radoman, J. (2018). *Small States in World Politics: State of the Art*. 179-200. 10.11643/issn.2217-995X182SPR101. DOI: 10.11643/issn.2217-995X182SPR101.

Rivera, J. (2015). Achieving cyberdeterrence and the ability of small states to hold large states at risk, in *2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace*, 2015, pp. 7-24, DOI: 10.1109/CYCON.2015.7158465.

Schneider, J., (2019). Deterrence in and through Cyberspace, in *Cross Domain Deterrence*, Lindsay & Gartzke eds. https://doi.org/10.1093/oso/9780190908645.003.0005.

Sweijs, T., Zilincik, S. Bekkers, F. & Meessen, R. (2021). *A Framework for Cross-Domain Strategies Against Hybrid Threats*, The Hague Centre for Strategic Studies, 2021.

Sweijs, T., Zilincik, S. (2021). The Essence of Cross-Domain Deterrence, in Osinga, F., Sweijs, T. (eds) *NL ARMS Netherlands Annual Review of Military Studies 2020.* NL ARMS. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-419-8_8.

Taddeo, M., (2018a). The Limits of Deterrence Theory in Cyberspace. *Philos. Technol.* 31, 339–355 (2018). https://doi.org/10.1007/s13347-017-0290-2.

Taddeo, M., (2018b). Deterrence and Norms to Foster Stability in Cyberspace. *Philos. Technol.* 31, 323–329 (2018). https://doi.org/10.1007/s13347-018-0328-0.

Toomse, R. (2015). *Defending Estonia in peace and war. Retaining a small state near aggressive neighbor by utilizing unconventional strategies.* Dissertations on social sciences, Tallinn University 2015.

Turvallisuuskomitea, (2013). *Suomen kyberturvallisuusstrategia*, ISBN: 978-951-25-2434-1 pdf.

Turvallisuuskomitea, (2019). *Suomen kyberturvallisuusstrategia 2019*, ISBN: 978-951-663-051-2 pdf.

Väyrynen, R. (1997). Small States: Persisting Despite Doubts in *The National Security of Small States in a Changing World*, edited by Efraim Inbar and Gabriel Sheffer, 41−77. London: Frank Cass.

Wanic, E. and Rowe, N. (2018). Assessing Deterrence Optinos for Cyber Weapons, in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 13-18, doi: 10.1109/CSCI46756.2018.00011.

Identifying Commonalities of Cyberattacks Against the Maritime Transportation System

Rebecca Rohan
Marymount University, Arlington, Virginia, United States of America

rebecca_rohan@marymount.edu

Abstract: The purpose of this study is to identify commonalities in cyberattacks against the civilian maritime transportation system (MTS). For this exploratory study, the researcher analysed documents to identify trends about the cyberattacks impacting and responsible adversaries targeting maritime operations. The MTS can use identified trends to make informed decisions about information technology (IT) and operational technology (OT) requiring new or enhanced cybersecurity measures. Current research examining publicly disclosed cyberattacks impacting MTS companies identifies the trend of increasing cyberattacks against the MTS. However, current research fails to examine adversaries and their social-political needs thoroughly. Knowledge of the adversary based on the Diamond Model of Intrusion Analysis can be augmented by identifying which MTS assets (e.g., shipbuilding, ports) and which aspect of the information security triad—Confidentiality, Integrity, or Availability (CIA)—the adversary targeted. At the conclusion of this limited, exploratory document analysis, the researcher determined the most compromised aspect of the information security triad was Availability and then Confidentiality; there were no identified Integrity compromises. The most targeted MTS assets was shipping companies, followed by ports, administration, shipbuilding, and vessels. Concerning the adversary customer behind MTS cyberattacks, China was first, followed by unknown cyber adversaries, then Russia, Iran, and Israel. Last, in terms of adversary's social-political needs, data exfiltration occurred the most, followed by ransomware, political agenda, and unknown needs.

Keywords: Maritime Transportation System, MTS, Adversary, Information Technology (IT), Operational Technology (OT)

1. Introduction

Over recent years, cybersecurity attacks have increasingly impacted the maritime transportation system (MTS), which plays a critical role in the global economy. The MTS moves over 80 percent of all goods around the globe (Statista, 2022). Cyberattacks, including deployment of malicious software (or malware), have disrupted operations at ports, in shipping companies, and in MTS administration (Dickerson, 2021; Vanguard, 2020). To better secure information technology (IT) and operational technology (OT), the MTS needs to understand better the adversaries behind the cyberattacks impacting maritime operations.

The purpose of this paper is to identify trends in adversaries targeting of the MTS by exploring cyberattacks impacting the MTS, including maritime administration, ports, shipbuilding, shipping companies, and vessels. In Section 2, related work pertaining to cyber adversaries per the Diamond Model of Intrusion Analysis and published research on cyber adversaries attacking the MTS will be reviewed. Section 3 will discuss the methodology and research design. Section 4 will present the results and discuss the findings. Finally, Section 5 will address the conclusion and recommendations for future work.

2. Related Work

Current research addresses concerns about increasing cyberattacks against the civilian MTS but does not provide an in-depth analysis of the adversaries behind those cyberattacks. The Maritime Transportation System Information Sharing and Analysis Center (MTS-ISAC) noted a 30 percent increase in reporting of malicious cyber activities against MTS stakeholders during January-June 2021 (Dickerson, 2021). Other works have examined factors affecting cybersecurity risks in the MTS without looking into the adversaries behind the cyberattacks (Tam and Jones, 2019). Current work has addressed implementing tools for dynamic risk assessments for improving cybersecurity in ports or managing cybersecurity risk using risk analysis models (Tam, Moara-Nkwe, and Jones, 2021; Paté-Cornell et al, 2018).

In the Diamond Model of Intrusion Analysis, the adversary compromises one of four points of the diamond with the three remaining points being victim, capability, and infrastructure. The adversary is an actor/organization employing a capability over an infrastructure against the victim to achieve an effect (Caltagirone, Pendergast, and Best, 2013).

Knowledge of the adversary is usually elusive and likely unknown when a cyberattack is initially identified (Caltagirone, Pendergast, and Best, 2013). Details about the adversary include the adversary operator, the adversary customer, and social-political needs for conducting a cyberattack (Caltagirone, Pendergast, and Best, 2013). The adversary operator is the actor/organization conducting the cyberattack, and the adversary customer

is the entity that will benefit from the cyberattack (Caltagirone, Pendergast, and Best, 2013). In conducting a cyberattack, the adversary can meet their social-political needs, or goals, of the attack (Caltagirone, Pendergast, and Best, 2013). In this study, the adversary's social-political needs fell into one of four categories: data exfiltration, political agenda, ransomware, or unknown. Knowledge of the adversary based upon the Diamond Model of Intrusion Analysis can be augmented by identifying which aspect of the information security triad—Confidentiality, Integrity, or Availability (CIA)—and which MTS asset (e.g., shipbuilding, ports) the adversary targeted during a cyberattack.

3. Methodology and Research Design

For the MTS, publicly available datasets concerning cyberattacks and the adversaries conducting said attacks are lacking. To complete this exploratory research, document analysis was based on two open-source websites:

- The Center for Strategic and International Studies (CSIS) maintains a Significant Cyber Incidents timeline from the year 2006 to the present focusing on cyberattacks against the defense sector, government agencies, high-tech companies, or those economic crimes in excess of one million U.S. dollars (Center, n.d.).
- The Council of Foreign Relations maintains a Cyber Operations tracker from the year 2005 to present involving publicly known state-sponsored cyber incidents (Council, 2022).

To focus on the entries pertaining to the MTS, the researcher:

1. Combined entries from CFR and CSIS into one spreadsheet for a total of 1,319 entries.
2. Deleted duplicate entries.
3. Deleted entries lacking the key terms *maritime, marine, transportation,* or *shipping* (at this point, 46 entries remained).
4. Deleted 20 entries with terminology pertaining to defense or military—*defense, military, naval technologies, navy, submarine, unclassified,* or *warship*.
5. Deleted 11 entries lacking specificity of a target in the MTS and lacking keywords *civilian maritime operations, maritime affairs, shipping,* or *transportation*.

See Figure 1 for a graphical representation noting how the CFR and CSIS entries were filtered.

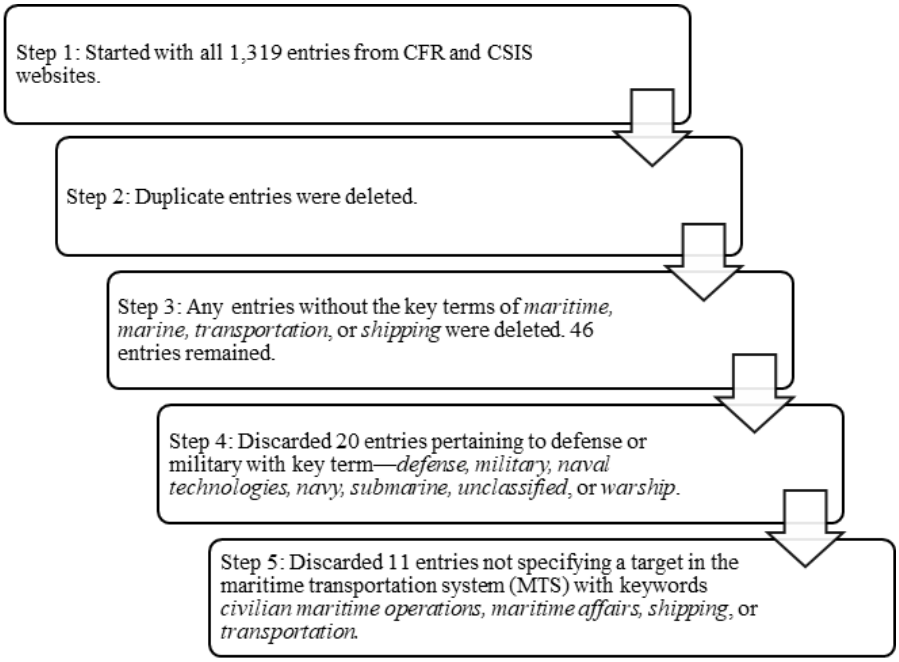


Figure 1: Steps for filtering cyber incident entries from CFR and CSIS.

4. Results and Discussion

Once filtering steps outlined in Section 3 and Figure 1 were completed, 15 entries remained of the original 1,319 entries obtained from CSIS' Significant Cyber Incidents and CFR's Cyber Operations Tracker (CSIS, 2021; CFR, 2021). The researcher created a spreadsheet with the 15 remaining entries to sort the data. The spreadsheet had the following columns:

- Year;
- Incident Description;
- MTS Asset:
    - Administration – maritime agencies or companies providing software used by the MTS;
    - Ports – ports or terminals where cargo is loaded/unloaded;
    - Shipbuilding – companies building ships used in the MTS;
    - Shipping Companies – companies handling the business or physical movement of cargo; or
    - Vessels – ships or merchant vessels transporting cargo.
- Victim of Cyberattack;
- Adversary Operator;
- Adversary Customer;
- Security Fundamental Compromised – Confidentiality, Integrity, or Availability; and
- Social-Political Needs – data exfiltration, political agenda, ransomware, or unknown need.

Arranged chronologically from newest to oldest, the resulting 15 entries on cyber incidents and adversary details are shown in Figure 2.

Year	Incident Description	MTS Asset	Victim of Cyberattack	Adversary Operator	Adversary Customer	Security Fundamental Compromised	Social-Political Need
2020	Iranian hackers compromised an Isareli developer of logistics management software and then accessed data over 40 clients of the developer (CSIS, 2021).	Administration	Unspecified	Unknown	Iran	Confidentiality	Political Agenda
2020	Iran's Ports and Maritime Organization experienced a cyberattack (CSIS, 2021).	Ports	Iran's Ports & Maritime Organization	Unknown	Unknown	Availability	Unknown
2020	The International Maritime Organization (IMO)--the United Nation's maritime agency--experienced a cyberattack impacting its website and networks (CSIS, 2021).	Administration	International Maritime Organization	Unknown	Unknown	Availability	Unknown
2020	CMA CGM SA, a French shipping company, experienced a ransomware attack in two of its Asian subsidiaries; the ransomware significantly disrupted information technology networks but did not impact movement of cargo (CSIS, 2021).	Shipping Companies	CMA CGM SA	Unknown	Unknown	Availability	Ransomware
2020	Operations at Iranian port, Shahid Rajaee, were disrupted for several days due to cyberattacks conducted by Isareli hackers (CFR, 2021; CSIS, 2021)	Ports	Shahid Rajaee Port	Unknown	Israel	Availability	Political Agenda
2019	A merchant vessel coming into the U.S. via interational waters reported a malware incident impacting the vessel's networks. The U.S. Coast received the report and issued a warning (CSIS, 2021).	Vessels	Unspecified	Unknown	Unknown	Availability	Unknown
2017	The Port of Rosario is a suspected victim of NotPetya malware that encrypted data on victim's computers without any means of data decryption (CFR, 2021).	Ports	Port of Rosario	Unknown	Russia	Availability	Ransomware
2017	For two days, Maersk's operations at port terminals were shut down by the NotPetya malware (CFR, 2021).	Ports	Maersk	Unknown	Russia	Availability	Ransomware
2017	For two days, Maersk's shipping operations were shut down by the NotPetya malware (CFR, 2021).	Shipping Companies	Maersk	Unknown	Russia	Availability	Ransomware
2015	OceanLotus has been conducting cyberattacks targeting shipping companies and marine agencies. Qihoo360, a Chinese company, reported OceanLotus has been running an espionage program since 2012 (CSIS, 2021).	Administration	Unspecified	Unknown	China	Confidentiality	Data Exfiltration
2015	OceanLotus has been conducting cyberattacks targeting shipping companies and marine agencies. Qihoo360, a Chinese company, reported OceanLotus has been running an espionage program since 2012 (CSIS, 2021).	Shipping Companies	Unspecified	Unknown	China	Confidentiality	Data Exfiltration
2015	Threat actor, Ocean Lotus, has conducted cyberattacks against maritime construction firms (CFR, 2021)	Shipbuilding	Unspecified	Unknown	China	Confidentiality	Data Exfiltration
2013	Threat actor, IceFog, conducts cyberattacks targeting maritime and shipbuilding groups. Targets have primarily been in Japan and South Korea (CFR, 2021).	Shipping Companies	Unspecified	Unknown	China	Confidentiality	Data Exfiltration
2012	Threat actor, Lucky Cat, conducts cyberattacks against Indian and Japanese shipping industries (CFR, 2021).	Shipping Companies	Unspecified	Unknown	China	Confidentiality	Data Exfiltration
2012	Threat actor, Sneaky Panda, targets cyberattacks against shipping firms in the private sector (CFR, 2021).	Shipping Companies	Unspecified	Unknown	China	Confidentiality	Data Exfiltration

Figure 2: Entries remaining after data filtering and sorted chronologically from newest to oldest

This exploratory document analysis on cyber incidents and associated adversary details are shown in Figures 3 through 6.

Figure 3 shows the cyber incidents per MTS asset:

- The most targeted MTS asset was shipping companies with six incidents.
- The second most targeted MTS asset was ports (four incidents) followed by administration (three incidents), shipbuilding (one incident), and vessels (one incident).

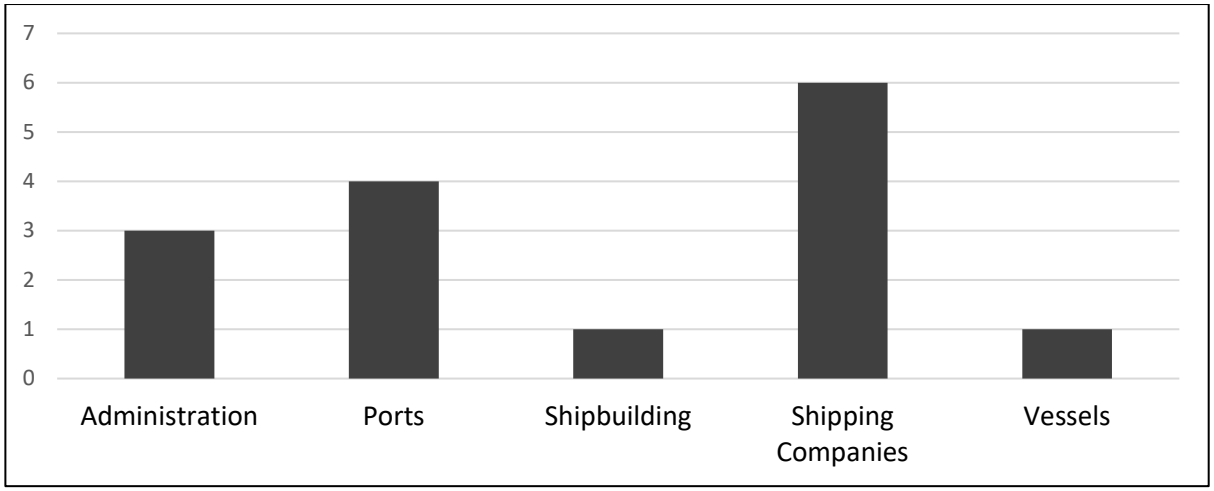


Figure 3: Cyber incidents per MTS asset

As shipping companies, ports, and administration assets incorporate more IT to support their operations, it follows these three MTS assets would be more often targeted in cyberattacks. The shipbuilding and vessels assets of the MTS have more OT deployed and may be harder to target, wittingly or unwittingly, during cyberattacks.

With respect to the adversary customers behind cyberattacks against the MTS assets, Figure 4 shows:

- China was the adversary customer for six of the cyberattacks—four in shipping companies, one in shipbuilding, and one in administration.
- Russia came in second as an adversary customer with three total cyberattacks, including two cyberattacks against ports and one cyberattack against shipping companies.
- Iran was an adversary customer for one cyberattack in administration.
- Israel was an adversary customer for one cyberattack in ports.
- Adversary customers were not known for one incident each in administration, ports, shipping companies, and vessels.

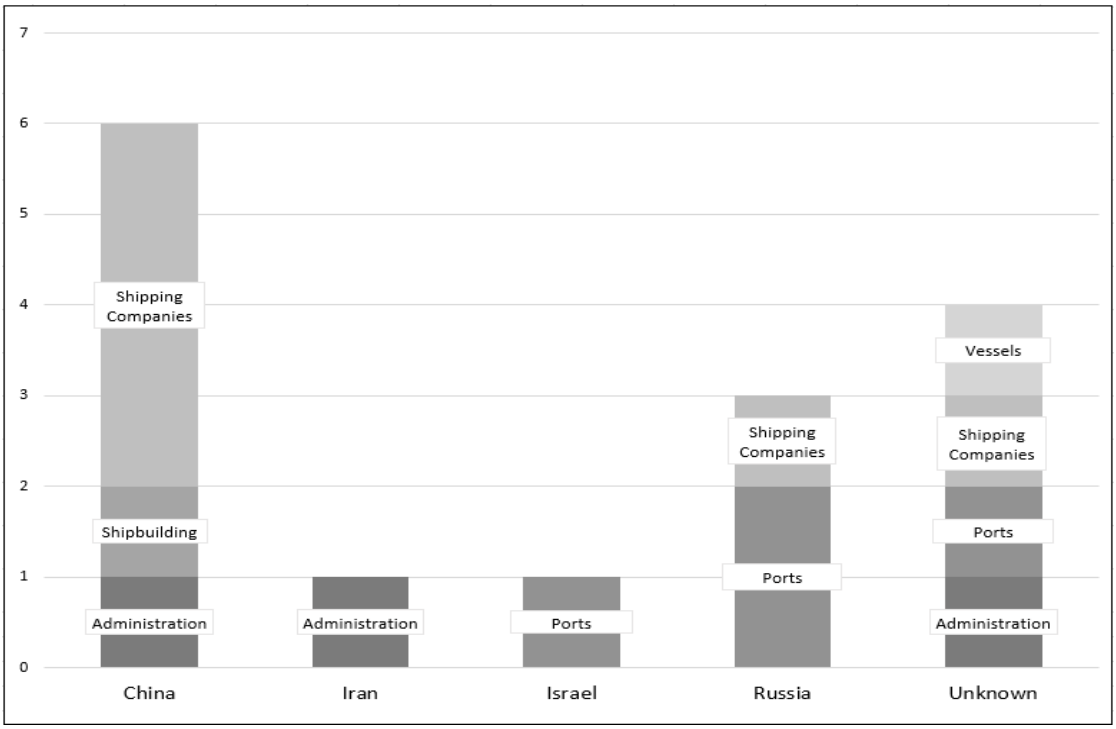


Figure 4: Adversary customer per MTS asset

Looking at the security fundamental compromised across the 15 cyber incidents, Figure 5 shows:

- Availability was compromised in eight of the incidents impacting every MTS asset except shipbuilding.
- Confidentiality was compromised in seven of the 15 cyber incidents impacting administration, shipbuilding, and shipping companies.
- Integrity was not compromised in any of the 15 cyber incidents studied.

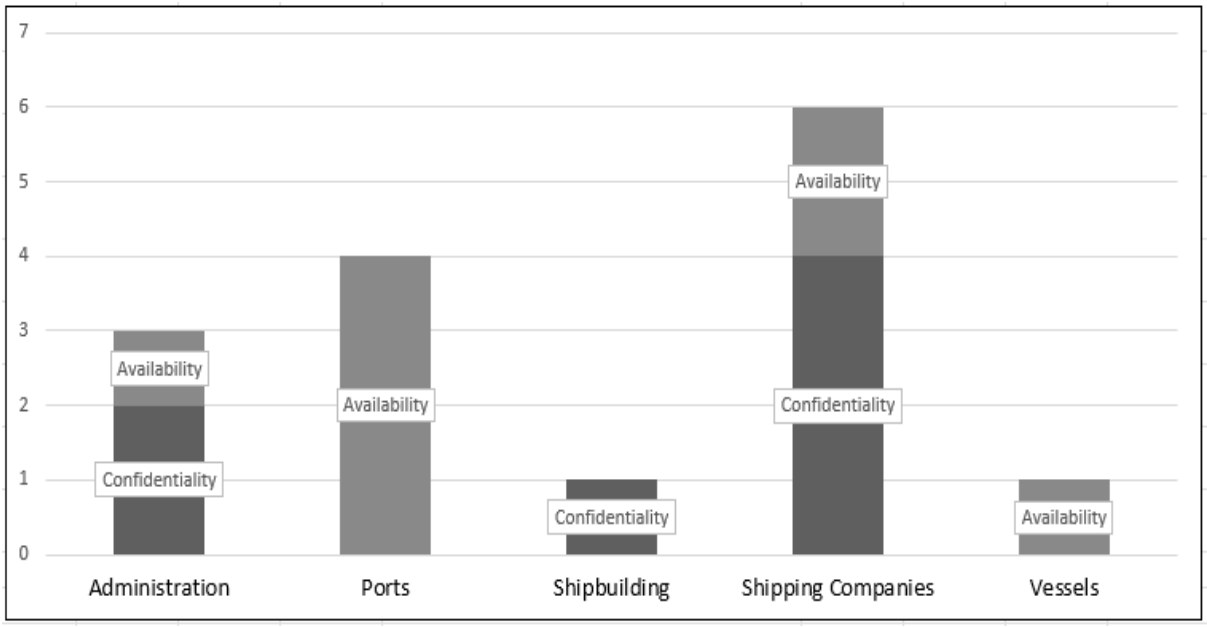


Figure 5: Security fundamental compromised per MTS asset

Last, the adversary's social-political needs per adversary customer, CIA triad aspect, and MTS asset are shown in Figure 6:

- Data exfiltration accounted for six of the cyber incidents in administration (one incident), shipbuilding (one incident), and shipping companies (four incidents).
    - China was the adversary customer for all six data exfiltration incidents.
    - Confidentiality was compromised in all six data exfiltration incidents.
- The second highest social-political need was ransomware at four incidents impacting ports (two incidents) and shipping companies (two incidents).
    - Russia was adversary customer for three ransomware incidents.
    - The adversary customer was unknown for one ransomware incident.
    - Availability was compromised in all four ransomware incidents.
- Political agenda had two incidents with one cyberattack each in administration and ports.
    - Iran was the adversary customer for one political agenda incident.
    - Israel was the adversary customer for one political agenda incident.
    - Confidentiality was compromised in both political agenda incidents.
- Last, three cyberattacks were categorized as unknown needs involving an unknown adversary customer with one incident each in administration, ports, and vessels.

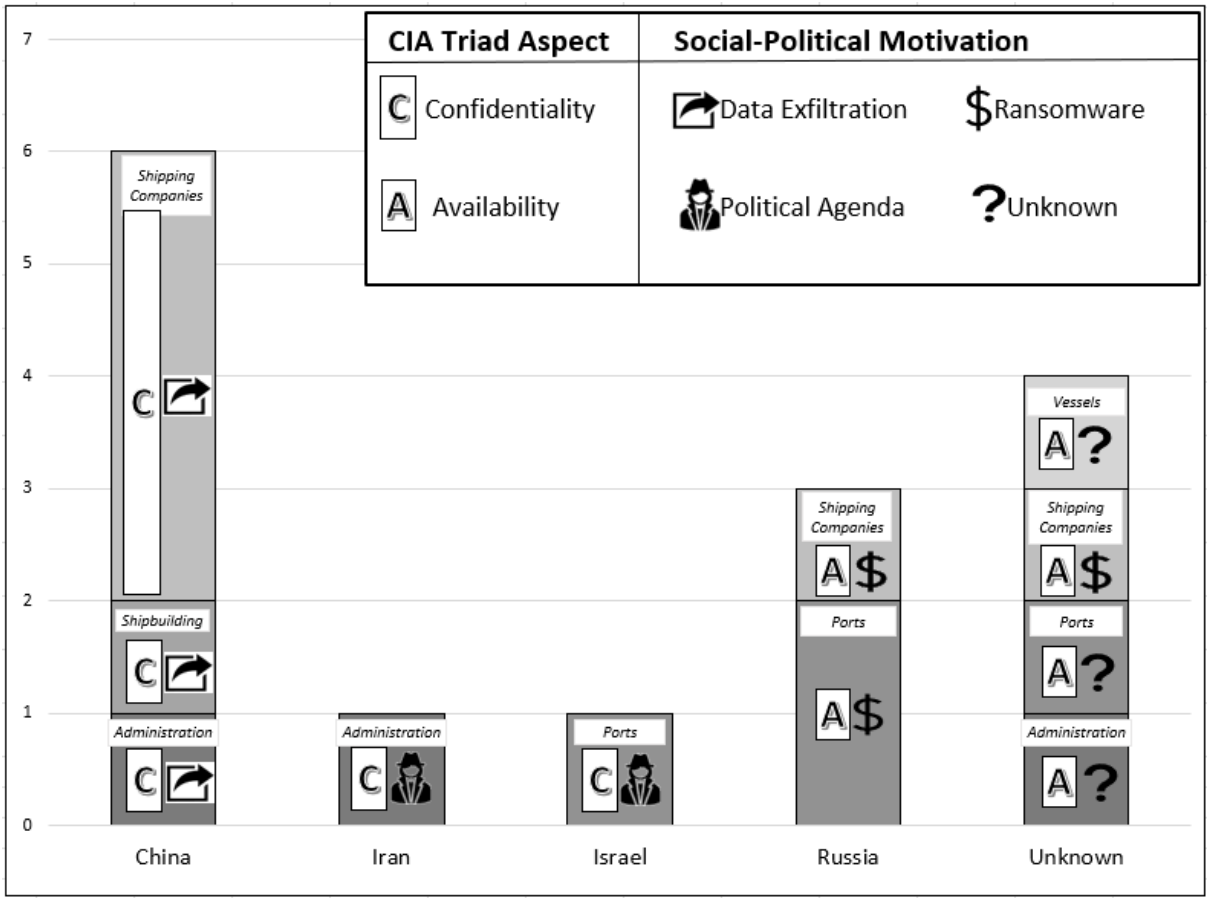


Figure 6: Adversary social-political needs per adversary customer, CIA triad aspect, and MTS asset

5. Conclusion and Future Work

Despite the lack of cyberattacks in publicly available data, this limited exploratory document analysis revealed that commonalities exist for adversaries conducting cyberattacks against the MTS. Data exfiltration and ransomware are concerns for administration, ports, shipbuilding, and shipping companies. Over half of the cyberattacks studied involved adversaries disrupting Availability, with the remaining cyberattacks impacting Confidentiality. China and Russia account for most of the adversary customers in MTS cyberattacks, but some adversary customers were unknown for a few cyberattacks.

Limitations identified during this exploratory document analysis of adversaries conducting cyberattacks against the MTS include:

- An overall lack of knowledge and research about adversaries conducting cyberattacks against the MTS;
- A lack of publicly available datasets for cyberattacks impacting the MTS;
- Information found in the two databases used, CSIS and CFR, provided some data about adversaries but additional details were found in open-source reporting;
- Cyberattack information concerning the MTS is disjointed and found across a variety of sources.

This work can be extended by including additional information from publicly available sources, especially sources focused on or popular across the MTS. Identifying amplifying resources could provide further data on adversaries, including adversary operators, adversary capabilities, and adversary infrastructure. Future research could address identifying the unknown adversary customers, unknown social-political needs, and the lack of information on adversary operators. In addition, future research could identify trends in adversaries' techniques, tactics, and procedures (TTPs) when conducting cyberattacks against MTS assets. Developing a database providing comprehensive data on cyberattacks against the MTS would be beneficial for organizations to identify and implement the appropriate security measures to harden IT and OT against cyber adversaries.

Acknowledgements

The author thanks Dr. Donna Schaeffer for her efforts as their project advisor. Additionally, the author thanks M.D. for reviewing and providing feedback on this paper.

References

Caltagirone, S., Pendergast, A., and Betz, C. (2013) *The Diamond Model of Intrusion Analysis* (Technical Report ADA586960), [online], Center for Cyber Threat Intelligence and Research, https://apps.dtic.mil/sti/pdfs/ADA586960.pdf.

Center for Strategic & International Studies. (n.d.) *Significant Cyber Incidents Since 2006*, [online], https://csis-website-prod.s3.amazonaws.com/s3fs-public/210430_Significant_Cyber_Events_List.pdf?B21zjJhsoO3qkgQNyGMmZN5IhAE80S_I.

Council on Foreign Relations. (2022) *Tracking State-Sponsored Cyberattacks Around the World*, [online], https://www.cfr.org/cyber-operations.

Dickerson, S. (2021) *The MTS-ISAC and NORMA Cyber Strengthen Information Sharing Ties*, [online], EIN News, https://www.einnews.com/pr_news/545608768/the-mts-isac-and-norma-cyber-strengthen-information-sharing-ties.

Paté-Cornell, M.-E., Kuypers, M., Smith, M., and Keller, P. (2018) "Cyber Risk Management for Critical Infrastructure: A Risk Analysis Model and Three Case Studies," *Risk Analysis: An International Journal* [online], Vol. 38, No. 2, February, pp 226–241, https://doi.org/10.1111/risa.12844.

Statista Research Department. (2022) *Container Shipping – Statistics & Facts,* [online], https://www.statista.com/topics/1367/container-shipping/.

Tam, K., and Jones, K. D. (2019) "Situational Awareness: Examining Factors That Affect Cyber-Risks in the Maritime Sector," *International Journal on Cyber Situational Awareness*, Vol. 4, No. 1, pp 40–68.

Tam, K., Moara-Nkwe, K., and Jones, K. (2021) "A Conceptual Cyber-Risk Assessment of Port Infrastructure," *2021 World of Shipping Portugal: An International Research Conference on Maritime Affairs [online]*, 28-29 January, https://pearl.plymouth.ac.uk/handle/10026.1/16704.

Vanguard News. (2020). *Maritime cyber attacks increase by 900% in three years*, [online], https://www.vanguardngr.com/2020/07/maritime-cyber-attacks-increase-by-900-in-three-years/.

Evaluating a Non-platform-specific OCR/NLP system to detect Online Grooming

Jake Street, Dr Funminiyi Olajide
Nottingham Trent University, Nottingham, United Kingdom

jake.street02@ntu.ac.uk
funminiyi.olajide@ntu.ac.uk

Abstract: Online Grooming is a social engineering attack in which the attacking party uses deceptive practices for sexual gratification. The targets of these attacks can vary in demographics however in most cases the target is children, with most of these attacks occurring on social media platforms. As well as the illegality of these attacks in the UK and US, children who experience these attacks are at a higher risk of self-harm or having suicidal thoughts. Due to the deployment of new social media platforms/features any implementation that is made specific to a certain feature/platform is likely to be outdated/ineffective upon release, due to the volatility of the methods/tactics used. Therefore a non-platform specific implementation has been considered within this investigation. From a preliminary analysis, it was concluded that there was an average true positive detection rate of 71% from using optical recognition and natural language processing across three different social media platforms. It is suggested that implementing this text extraction and processing method alongside a 'category-based' machine learning algorithm, a solution that can identify online grooming can be developed that considers the 'real world complexities' of this attack.

Keywords: Natural Language Processing, Machine Learning, Optical Character Recognition, Online Grooming

1. Introduction

Individuals that carry out social engineering attacks have a wide range of motives as to why they are carrying out their attack. Some social engineers have similar motives to attackers that would take a more technical-based approach, with these attacks often being targeted at commercial organisations. The motives for conducting these attacks are likely to be for financial gain, to disrupt and cause harm to the organisation (possibly a disgruntled employee), or a competitor of the organisation aiming to gain a competitive advantage.

These social engineering attacks targeted at organisations follow the hypothesised 'CIA triad' of information security (Samonas, 2014) in which the 'CIA triad' refers to Confidentiality, the principle in which only authorised entities can access a given piece of information; Integrity, the principle in which information is not modified without proper authorisation; and Availability, the principle in which information is available upon demand by an authorised entity.

Table 1 demonstrates the different social engineering attacks targeted at organisations that specifically relate to each of the principles of the 'CIA triad'.

Table 1: Social Engineering Attack Methods that Affect each Element of the 'CIA triad'

'CIA triad' Principle	Social Engineering Attack Method
Confidentiality	Social biases/asking questions
Integrity	Requests while impersonating an authority figure
Availability	Theft of infrastructure (to affect a service)

However with social engineering attacks unlike most technical attacks the target that is selected determines the likelihood that the attack will be successful, the likelihood that the attack will be identified by the victim, and the severity (risk) to the attacker if the attack were to be identified by the victim.

Because of these factors victim selection is one of the main considerations for an attacker when conducting one of these attacks, with attackers looking for specific personal attributes when selecting their target or evaluating the likelihood of success/risk of the attack.

Alike a vulnerable computer system, individuals that have a high attack success rate based upon their attributes should be referred to as 'Vulnerable individuals' with each of these vulnerabilities often having their own

corresponding exploit/attack. Some potential social engineering exploits/attacks for each of these personal vulnerabilities are shown in Table 2.

Table 2: Personal Attributes with a Perceived Increased Attack Success Rate

Vulnerability/Personal Attribute	Social Engineering Attack
Naivety	All Attacks
Fear/Anxiety	Tech support scams
Technical Awareness	Tech support scams, Phishing emails
Lack of Awareness of Sexual Exploitation	Online Grooming, Relationship Scams

1.1 Context

This investigation will be observing the Online Grooming social engineering attack which is defined as the process in which an adult forms a sexually abusive relationship with a child using technology (Lorenzo-dus, 2017). These attacks have been found to be increasing in frequency over time (Bentley, 2018) despite an increased focus on education for children surrounding this (Department for Education, 2019) in the UK. Therefore due to this perceived lack of effectiveness of education to mitigate these attacks, a greater focus on technical solutions should be considered.

2. Literature Review

Online Grooming from a cybersecurity approach is in its infancy within Literature, due to Online Grooming being predominantly seen within scope of the domain of Psychology and Social Sciences as opposed to Cybersecurity and Social Engineering. This should be considered when analysing this literature so that the findings presented can be best considered for a cybersecurity approach.

2.1 Psychological Considerations

There are many different methods/strategies that are used by online groomers when conducting their attack which makes creating an effective solution technically complex. Despite this variety of approaches taken by attackers, there remains constants/'phases' between all OG attacks that perpetrators tend to follow throughout communication with the target. These phases are based on a chatroom context so therefore it should be noted that these phases are likely not generalizable to other contexts/mediums that are used within OG attacks (e.g. within a social media live streaming service).

Initiation Phase: The purpose of this phase is to determine the suitability of the target. A suitable target will depend on the attacker's preference in terms of age and gender. This is often the first communication between the attacker and the target which is done to scope the target for personal details.

Determining Method Success. Once the target has been deemed suitable, the groomer may or may not attempt to ask questions to ascertain what method would be best to execute the attack. This commonly involves having 'friendship-forming' conversations, which is often used to identify what the target is likely to respond positively to (O'Connell, 2003).

Determining Attack Risk Phase: This is based upon the groomer's risk tolerance they may or may not 'complete a risk assessment' to determine the risk that they are likely to face from executing the attack (O'Connell, 2003). This risk assessment shall be used to determine the probability of getting detected as well as the severity of detection.

Sexual Gratification Phase: The final stage of the grooming process is executing the attack, often with the intent of sexual gratification for the attacker. This can include requests from the attacker or more manipulative strategies such as blackmail.

2.1.1 Victim-centric Approach

While many victims of Online Grooming and sexual abuse feel it is their fault for the attack occurring (Patel, 2021), this is clearly not the case. However, there are a variety of attributes and behaviours that appear at a higher frequency for individuals who have experienced one of these attacks as opposed to the average person.

It should be noted that without a longitudinal study of children throughout growing up, it cannot be concluded that these attributes/behaviours have an impact on the chance of an attack occurring or if these attributes/behaviours are developed from experiencing the attack in the first place. This latter point does have some validity, as Patel (2021) observed that individuals that have experienced this form of abuse are more likely to seek similar abusive scenarios in a process known as 'revictimization'.

'Vulnerable behaviours' in this context refers to activities that a child does which will increase the probability of an attack occurring and/or activities that increase the severity of an attack. For example, a vulnerable behaviour that increases the probability of an attack occurring would be going on chatroom or other 'person discovery' websites as 10-20% of children that visit these websites receiving sexual messages from adults (Finkelhor, 2000).

An activity deemed a 'Vulnerable behaviour' that increases the severity of the attack would be a child using a livestream feature within a social media platform that allows comments from watchers. This increases the severity of a potential attack as the child is put under a time constraint to respond to comments from watchers as well as the potential for multiple attackers to gang up on one target.

2.1.2 Perpetrator-centric Approach

It is suggested that some perpetrators, which O'Connell (2003) categorised as "cyber rapists", are not concerned about mitigating their risk, by carrying out a 'risk assessment'. Instead, it is suggested that a 'hit and run' tactic is used, which involves outright aggressive acts towards the victim. This strategy aims to capitalise on the fear of the victim, as opposed to a strategy in which the victim is given more control. These attackers rely on any possible anonymity they may have and constant threats to the victim. It is assumed from the description given by O'Connell (2003) that when an attacker deems a situation as too risky, they attempt to mitigate the attack by "running" and ceasing all communication, as well as possibly using destructive measures or further threats to the victim which can be attributed to the 'Damage Limitation' phase (O'Connell, 2003).

In pretty much all cases, grooming is done purely for sexual gratification. However as deemed by Wood (2013) paedophiles do not feel "wholly adult", so it could be hypothesised that some grooming attacks could be done for social reasons.

Conte (1989) discovered that most (19 in 20) groomers had a "secondary deviance", in addition to their main grooming behaviour. Therefore, it can be likely that these may be combined in their attacks, and thus should be considered when designing a solution to detect these attacks.

2.2 Current Solutions

This section shall describe the solutions that have been implemented to attempt to mitigate Online Grooming attacks. Some of these implementations may be more widespread than others, with some solutions described here being hypothetical.

It should be noted that these solutions are likely to differ in appropriateness based on if they are to be implemented within a school or home setting: with a school policy being more focused on security and well defined, and with a home policy being more focused on usability and is often undefined.

2.2.1 Website Blocking (Allow list & Deny list)

This solution involves stopping a Web browser from accessing a list of websites (in terms of the 'Deny list' strategy) or only allowing the Web browser to access a list of websites (in terms of the 'Allow list' strategy). These can be implemented to stop a child from accessing social media or other platforms that paedophiles tend to operate on.

The issue with implementing this is that the child is likely to not comply with these restrictions and it is easy for the child to get round these restrictions as they work at browser level (so use of a different browser/device would circumvent this mitigation method) or by finding a communication platform that has not been denied by the list.

2.2.2 Activity Monitoring

Activity Monitoring refers to the recording information about the utilisation of software on a system and the specific activities carried out on the software. Attributes that tend to be recorded within this include duration of an activity, files downloaded, messages sent and received, and a timestamp of the activity (Netnanny.com,

2022). This information can be used to detect if an Online Grooming attack is taking place however there are some significant limitations to this. Firstly, if there is no further automation, for a parent/guardian to be able to detect an Online Grooming attack it would take a significant amount of human processing, as each activity would need to be looked through in detail, which would be infeasible for most guardians. Secondly, for this implementation to be effective at helping a parent/guardian detect an online grooming attack, this is likely to be deemed an invasion of the child's privacy and could aid an abusive parent/guardian. This is likely to also cause the child to not comply with the system and look for methods to circumvent it.

2.2.3 Word Blocking

Word Blocking refers to software that can parse on-screen text on specific websites and compare this text to a list of dangerous words/phrases that refer to a variety of topics that might not be suitable for children. Following a detection of one of these dangerous words/phrases a variety of actions can occur, but most commonly one of the following; the child will not be allowed to access the website, the words/phrases are removed from the html (client-side), or the child will be able to access the website but the parent would be informed of the threat (via a screenshot of the threat or relaying the text in question).

While this implementation can allow for the real-time detection and/or blocking of a threat, this solution suffers from the general complexities of being able to detect online grooming attacks (namely being able to establish the context of a communication). This solution also suffers in terms of being functional across multiple websites, as each platform is likely to require some fine tuning with the text that should be analysed by the software (to reduce the frequency of false positives).

2.2.4 Policy Publication

Policy publication refers to the non-technical mitigation method in which users on a given network/device/service agree to follow a specific set of rules. These policies are commonly implemented in schools (Cramer, 2010) as schools tend to have a clear definition as to what content/services are appropriate for their students to access and what is not. However, in most cases, this is not the same within the home environment due to the relaxation of restrictions as a child gets older e.g. a child being able to create a social media account when turning thirteen. It should be noted that despite these service-level policies, many children will use a service when they are not of appropriate age (Office of Communications, 2017).

While these policies do not detect or stop an attack from occurring, these can stop children engaging in vulnerable behaviours on networks/devices with these policies applied. The success of these policies are based on the mutual respect of the parties involved in the policy, as if one party feels that the policy is unfair upon write up or if a party does not comply to the rules of the policy then it is not likely to be respected by all parties.

2.2.5 Education Methods

Education methods refer to the process of educating both children (Department for Education, 2019) and their parent/guardian on vulnerable behaviours and how to identify online grooming (from both the child's and parent's perspective). While parents tend to be aware of Online Grooming as an issue that affects children, children tend to not be aware of the complexities of why adults engage in paedophilic behaviour and the risks that this poses to them.

Without providing children with detailed information about these attacks, which would raise significant ethical concerns, it is difficult to rely on this method to ensure that children will firstly identify the attack and secondly know the risks that face them because of it.

3. Design

3.1 Design Scope

Within creating a solution to detect online grooming attacks, a variety of factors need to be considered to ensure that an appropriate solution is made for a specific real-world use case. The use case in question for this investigation is for a solution that will be implemented within the home environment, with the assumption that the websites/platforms do not allow direct text-extraction from an API or the website's HTML. Additionally the child shall only access these websites through a desktop computer (as opposed to a phone or tablet devices).

It is important to remember the real-world context of an implementation of this nature, as there are many ethical challenges within this area. A significant challenge with implementing an online grooming detection system is protecting children from abuse from the parent/guardian. This is because, if a 'screenshot' feature is implemented within the solution, a parent/guardian alongside the attacker may be able to blackmail the child. Additionally, the parent may find out information about their child which could cause harm e.g. the system exposing the sexuality of the child. Another real-world consideration to make is the use of 'blocking' when a detection is made, to attempt stop an attack occurring in real time, however the issue with this is that it is likely to annoy users if there is a false positive and will therefore compromise the usability of the system. These factors must be considered within the solution, Figure 1 demonstrate the 'trade offs' that need to be considered when designing a software solution for online grooming.

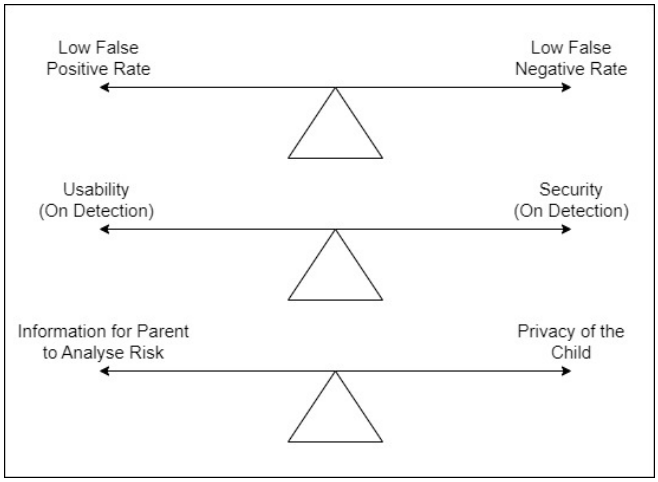


Figure 1: Design 'trade-offs' for Online Grooming Software Solutions

3.2 Design

Based upon the analysis of the current solutions it is deemed that implementing a 'Word Blocking Regular Expression' solution would be most effective considering the design scope. This will allow for some level of Online Grooming phrase detection to occur while being able to maintain a good level of privacy for the child (as opposed to using an 'Activity Monitoring' solution).

A design consideration that should be made is if the system should incorporate networking as there are some benefits in terms of features that could be implemented, these considerations are shown within Table 3.

Table 3: Advantages between an Offline and Networked Design

Advantages of an offline solution	Advantages of a networked solution
Does not require an internet connection to be functional.	Parents can access information from anywhere with Internet access.
Does not affect Internet speed, for other services/applications.	Ability to implement notifications for the parent
	Ability to instantly implement an updated word/phrase blocking list for all users.

Another complexity that needs to be considered within this use case is how to extract the text from the websites if an API or HTML parsing cannot be used. To be able to do this, screenshots of the current activity at a set regular interval in combination with an Optical Character Recognition (OCR) module shall be used. By processing the screenshots of the online activities done through OCR a string of text will be outputted.

This string of text will then be processed to ensure that any artefacts within the string, based upon misinterpreted features of the screenshot, are removed. Then following this these words will follow a splitting process to allow for further processing and for an ultimate decision on if the given image contains Online Grooming material. The individual words will go through Stemming and Lemmatization, this is to ensure that

phrases that use slightly different wording are normalised to reduce the number of false negatives (e.g. “word”, “words”, “worded”, and “wording” would all be Stemmed to “word”). Following this the individual words and phrases can be compared against the blocked words and phrases and then if a match is made a given action can be performed based upon the given use case (this can range from ‘locking’ the computer to sending a notification to the parent of a potential threat). This process is detailed in Figure 2.

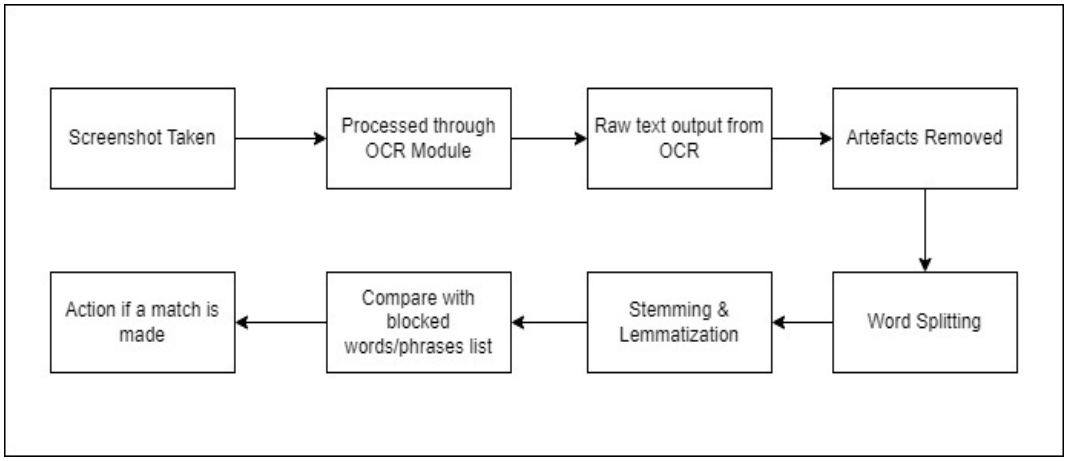


Figure 2: High-level Description of Solution

4. Methodology

The research philosophy behind this investigation is a proof of concept study to identify if using OCR is a feasible technique for text extraction for the purpose of identifying Online Grooming attacks over multiple social media platforms, or if ‘direct’ text extraction methods (such as APIs and HTML parsing) should be used.

Within this study only one OCR piece of software shall be evaluated, further research should investigate different OCR solutions to identify if there is a significant difference in text extraction for screenshots of websites.

In addition, this investigation shall only consider one screen size (23 inch). Other screen sizes should be investigate in further research to identify how this impacts accuracy, as well as observing how ‘mobile versions’ of websites affect this accuracy.

This system shall be tested using a random sample of 40 chatroom transcripts that were obtained from a ’honeypot’ Online Grooming study in the United States , sourced from ChatCoder.com (2022).

The testing of this system shall be conducted over three social media platforms: ‘Facebook’ from Meta Platforms; ‘Twitter’; and ‘e-Chat’, an anonymous unmoderated chatroom service where users can set up their own chatrooms.

Messages from ChatCoder.com (2022) will be sent from the ‘Attacker’s Computer’ to the ‘Child’s Computer’, on a variety of social media platforms, which shall be running the detection software described in Figure 2.

The Null hypothesis that shall be tested within this investigation is as follows: “There is no significance in the proportion of true positive detections to false negative detections while using OCR text extraction with NLP analysis for a specific social media platform”.

5. Results

Table 4 shows the true positive rate, false negative rate, and if the test null hypothesis is accepted or rejected for each of the social media platforms tested. As well as a collated, overall, result over all tests.

Within this investigation a ‘Correct Detection’ refers to an action, in this case redirecting the child’s PC to a certain website, occurring within a period of 30 seconds of a message that should be blocked by the system being sent to the child’s PC. Conversely, a ‘Missed Detection’ refers to the action occurring greater than 30 seconds from when the messages was sent or if the action did not occur from the message being sent.

Table 4: Investigation Results

Test Conducted	Correct Detections (True Positive rate)	Missed Detections (False Negative rate)	Outcome
Twitter	18	12	Accept Null
Facebook	20	10	Accept Null
e-Chat	26	4	Reject Null ($p < 0.05$)
All	64	26	Reject Null ($p < 0.05$)

These results show that out of the three social media platforms that was tested, only one test ("e-Chat") rejected the Null hypothesis ($p < 0.05$). The collated results across all three tests also rejected the Null hypothesis ($p < 0.05$).

6. Discussion

From these results it can be concluded that there is some significance in true positive detection using OCR, and while this result is somewhat promising some considerations should be made surrounding it.

Firstly, despite the statistical significance within the 'e-Chat' and 'All' tests the level of significance was low compared to an acceptable level for this use case. This is because there are clearly repeatable issues with using OCR text extraction in this context as even on a website that has a good density of text vs. images/other media, such as 'e-Chat', there were still some frequency of Missed Detections.

It can be assumed that the frequency of Missed Detections from text extracted through bespoke-to-service HTML parsing is close to zero, therefore this frequency of Missed Detection is too excessive for this use case within this implementation.

Another element that needs to be considered within the real-world implementation of an OCR/NLP solution to Online Grooming is the False Positive rate. This was not considered within this investigation, as it was out of scope, however it is likely that this will cause some usability impact on users if this were to be implemented at current.

Overall, it is suggested that further research is conducted on the effectiveness of an OCR/NLP solution on mobile devices as it is hypothesised that the results obtained would have greater ratio of true positives vs. false positives compared to the ratio obtained in this study due to less potential OCR artefacts within the text capture, reduced processing times due to smaller images, and greater OCR conversion accuracy due to an increased text size (in relation to the screen/image size).

References

Bentley, B. C. G. G. H. L. M. O. P. P. S. S. V. W., 2018. How Safe Are Our Children: National Society for the Prevention of Cruelty to Children. [online] Available at: <https://careleaverpp.org/wp-content/uploads/2018/07/how-safe-children-2018-report.pdf> [Accessed 9 October 2022].

Chatcoder.com. 2022. ChatCoder HomePage. [online] Available at: <https://chatcoder.com/index.html> [Accessed 15 October 2022].

Conte, J.R., Wolf, S. And Smith, T., 1989. What Sexual Offenders Tell Us About Prevention Strategies. Child Abuse & Neglect, 13(2), pp.293-301.

Cramer, M. and Hayes, G., 2010. Acceptable use of technology in schools: Risks, policies, and promises. IEEE Pervasive Computing, 9(3), pp.37-44.

Department for Education, 2022. Teaching Online Safety in Schools. [online] Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811796/Teaching_online_safety_in_school.pdf> [Accessed 9 October 2022].

Finkelhor, D., Mitchell, K.j. And Wolak, J., 2000. Online Victimization: a Report on the Nation's Youth.

Lorenzo-Dus, N. and Izura, C., 2017. " cause ur special ": Understanding trust and complimenting behaviour in online grooming discourse. Journal of Pragmatics, 112, pp.68-82.

Netnanny.com. 2022. [online] Available at: <https://www.netnanny.com/> [Accessed 15 October 2022].

O'Connell, 2003. A Typology of Child Cybersexploitation and Online Grooming Practices. [Online] Available at: http://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf [Accessed 13 October 2022].

Office of Communications, 2017. Children and Parents: Media Use and Attitudes Report. [online] Available at: <https://www.ofcom.org.uk/__data/assets/pdf_file/0020/108182/children-parents-media-use-attitudes-2017.pdf> [Accessed 7 October 2022].

Patel, P., 2021. Tackling Child Sexual Abuse Strategy 2021. [online] Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973236/Tackling_Child_Sexual_Abuse_Strategy_2021.pdf> [Accessed 9 October 2022].

Samonas, S. and Coss, D., 2014. The CIA strikes back: Redefining confidentiality, integrity and availability in security. Journal of Information System Security, 10(3).

Masters Research Papers

Demonstrating Redundancy Advantages of a Three-Channel Communication Protocol

Scott Culbreth and Scott Graham
Air Force Institute of Technology, Wright-Patterson Air Force Base, United States

Scott.Culbreth.1@us.af.mil
Scott.Graham@afit.edu

Abstract: Multi-Channel communications have the potential to provide advantages in security and redundancy. One widespread example of additional security is the use of 2 Factor Authentication wherein an authorization code is sent via a separate channel. As another example, spread spectrum technology offers resilience against channel interference. However, no currently deployed communication protocols take advantage of the full spectrum of security and performance gains that can be obtained through transmitting data over multiple channels. Taking inspiration from Redundant Array of Inexpensive Disks (RAID) and their use of data striping and mirroring, a secure multi-channel communication protocol was developed that is able to have greater security than an equivalent single channel system while also having greater resiliency against data corruption, therefore requiring fewer, if any, retransmissions than a single channel system when operating in a low availability environment. This approach admittedly comes with significant overhead, both in the use of additional channels and the need for additional processing. Whether the security and availability gains are worth the costs is an open question, with specific answers highly dependent on the particular applications. A specific multi-channel communication protocol, with three independent channels, and incorporating duplication on the bit level, was built and exercised within the OMNeT++ simulation environment in order to examine specific aspects of performance and security of the protocol. This exercise demonstrates that a secure multi-channel protocol operates with less latency than an equivalent single channel system when experiencing less than 50% channel corruption due to adversarial injection, resulting in reduced data loss and need for re-transmissions.

Keywords: Multi-Channel, Security, Encryption, Networking

1. Introduction

As threats to computer security evolve, our methods of combating security threats must also evolve. Direct threats to the confidentiality, integrity, and availability of communications are developed daily; this research seeks to develop an effective security-focused communication protocol operating at the session layer, taking advantage of diverse multi-channel communications. The proof of concept for this protocol includes a three-channel system that sends individual bits of data over two separate channels, which is referred to as a duplication factor of two (Simon 2021). The implementation incorporates bit scrambling, error detection and error correction in an attempt to address confidentiality, integrity, and availability of the system. The protocol was implemented and tested in simulation to demonstrate its effectiveness, and test the limitations. Simulation demonstrated improved performance regarding confidentiality, integrity, and availability in non-permissive environments without using encryption.

1.1 Background

Computer security is often represented using the Confidentiality, Integrity, and Availability (CIA) triad. These three factors often guide the actions taken to show that a computer network is secure and robust. Threats to confidentiality are those that allow an adversary to glean unintended information from what was intended to be a secret communication. This is often the primary area considered in "security" as it is the protection of secret information. Confidentiality threats include decrypting secret messages, but data leakage is also a major concern. Integrity ensures that the message received is the message that was sent, i.e., that data is not lost or manipulated in transit, either through random noise within the transmission medium, or through adversaries suppressing or injecting information. Finally, availability is a key consideration for security. If a service is unavailable, it does not matter if it is unreadable to an adversary as it is unusable by the intended parties, thus a system must be available to be viable.

1.2 Problem Statement

The desire for computer communication protocols to be fast, secure, reliable and inexpensive is obvious, but these objectives present a trade-off. Often a fast and secure system is not reliable and affordable while a fast and affordable system is generally less secure or reliable. A single channel system that is fast, secure, and reliable

is unlikely to be affordable. This research tackles this issue with a security focused multi-channel protocol. The protocol is focused on data leakage that could occur across the channels, or data an eavesdropper could gain by listening in on the channel and sniffing packets. In a single channel setup, all the data is available to a potential eavesdropper and all security would have to be based on cryptographic means. Encryption algorithms can carry significant overhead and can in some cases be unrealistic to use. Considerations for potential data injection attacks should be made as well. Finally, although networks typically offer multiple paths to a destination, most routing only takes advantage of a single path.

1.3 Hypothesis

A secure multi-channel protocol with a duplication factor of two can provide higher availability, with less data leakage, over an equivalent single channel system. Additionally, under heavy levels of corruption, a redundant multi-channel communication system has the ability to get messages through in less total time or with less latency per message/packet.

2. Related Work

This section provides a summary of existing networking technologies, how they inspired the creation of the protocol, and tests of the networking protocol. These include routing, data fragmentation and scrambling, as well as general network security requirements and background. Additionally, the simulation environment will be discussed, providing context on how the tests and exercises were conducted.

2.1 Multi-Channel Use

The use of multiple channels has existed in various forms for decades. The Public Switched Telephone Network adopted Signalling System 7 (SS7) in 1989, with both a communication channel, as well as an out of band signalling and control channel to operate and administer the network. This allowed for call data to operate on one channel while not being in competition with necessary signals to operate the network (Modarressi 1990).

Redundant Array of Inexpensive Disk (RAID) arrays can also be considered a multi-channel network with each disk acting as a channel. The methods used to duplicate and stripe data for redundancy, efficiency, and performance can be applied to other multi-channel uses (Hughes 2005).

Finally, Two-Factor Authentication (2FA), is another example of multi-channel use, and in this case, specifically for security. 2FA uses multiple channels to confirm the identity of a person attempting to access a system or account, beyond the traditional username and password, by sending an identifying code through a separate channel. This extra requirement helps ensure that the person providing login credentials has authorized access, as compromising both the username and password, as well as the secondary channel is significantly difficult (Dmitrienko 2014). Typically, this confirmation comes in the form of a text message or call to a mobile device, but various schemes exist. For example, the use of a secondary smart card provided in combination with login credentials or biometrics can be considered 2FA. Overall, the use of multiple channels creates significant obstacles for any adversary to overcome and results in significantly greater security (Dmitrienko 2014).

2.2 Multipath Transmission Control Protocol (MPTCP)

MPTCP is an extension of the traditional Transmission Control Protocol (TCP), utilizing multiple paths to improve performance and redundancy. Traditional TCP is limited to communication between two separate addresses, or interfaces, on a single channel. MPTCP extends TCP from communication between two interfaces to communication between two hosts or endpoints (Ford 2020). This enables multiple different interfaces from a single host to establish individual TCP connections and utilize both simultaneously. A typical example of this in practice would be a smart phone that has both WiFi and cellular connection capability. In a standard TCP connection, these two interfaces cannot be used in conjunction. However, when utilizing MPTCP, both the WiFi and cellular interfaces can be used simultaneously and thus allow for multiple paths to be used, allowing traffic to be sent over either of the two (Bonaventure 2016). If, for example, the user loses the WiFi connection, traffic is sent over the cellular connection alone, transparent to the user. Overall, this creates a protocol that is capable of extending a widely adopted protocol in the form of TCP and using multiple paths to improve redundancy.

2.3 OMNeT++

OMNeT++ is "an extensible, modular, component-based C++ simulation library and framework, primarily for building network simulators." (What is OMNeT++? 2022). This framework allows the creation, testing, and development of network characteristics within a Graphical User Interface supported, C++ based, environment. OMNeT++ has countless pre-built modules from which simple networks or queues can be modelled, or complex networking relationships can be analysed. Features such as Peer-to-Peer networks, wired and wireless connectivity, and cloud-based solution modelling are available. Additionally, due to the C++ based nature of the environment, any specific modules for a particular scenario can be created and integrated with the existing models and packages. This creates a seamless and easy to work with simulation environment specifically tailored to the work being done.

3. Methodology

The intent of this proposed protocol is to increase security by utilizing multiple channels and bit scrambling while also increasing availability by overcoming packet loss and corruption/injection through the multi-channel setup. Additionally, performance should also improve over a single channel system. We begin by defining requirements for the protocol, then developing to those requirements, followed by identifying metrics to quantify the protocol's performance and finally exercising the protocol in simulation to determine performance and results.

3.1 Defining Requirements

Defining the requirements for the protocol is essential to guide the development and implementation of the protocol. For this implementation, the protocol will have three channels. As described in Section 2, multiple channels are not a new concept, however the spreading of data across multiple channels for security benefits is novel (Simon 2021).

The next requirement is that of a duplication factor of 2 (DF = 2). This means that each bit of data is sent over two of the three available channels. This allows for redundancy; as corruption or suppression of data occur, the data can still reach the destination over another channel.

An additional requirement is for the protocol to be able to detect corrupted packets due to random bit flips or from malicious action (injection). The protocol should also implement some form of data scrambling. Data scrambling is the process through which individual bits are taken from original data and, as the name implies, scrambles them into different locations within the packet. An eavesdropper who is not privy to the scrambling code would find the data unreadable. In addition, the bits are spread across channels, such that no channel has the full set of bits, which we call "data dropout", creating a significant challenge for eavesdroppers looking to extract usable data within a stream (Bao 2017). While not as secure as encryption, scrambling combined with data dropout significantly increases attacker difficulty as the target data is now out of order and incomplete.

3.2 Developing the Protocol

The proposed protocol assumes operation at the session layer. The first challenge of a multi-channel setup is creating the mechanisms which handle the decision making and processing of packets into and out of streams. The initial setup and protocol design developed in (Simon 2022) is followed in this new protocol.

The *Dissembler* and *Assembler* work in pairs at the session layer to send and receive the packets respectively. The *Dissembler* distributes the data from the application and creates the *Channel Packets* which contain the data sent along the respective channel, along with all necessary headers and footers for the protocol to operate. This operates similar to the MPTCP endpoints described within Section 2.2 where the (*Assembler Dissembler*) pair are the endpoints that create the multiple paths. The primary difference is the security elements implemented within the *Dissembler* and the distribution of data across the channels.

The *Assembler* receives packets from each channel and populates the *Forwarding Buffer*, which contains the reassembled (unscrambled) data to be forwarded to the receiving application.

Figure 1 shows the structure of the network showing the *Dissembler* on the left, followed by an arbitrary number of intermediary nodes, representing other hops along the network. Intermediary nodes are abstracted from an arbitrary number of hops into a single node for purposes of simulation. The intermediary nodes act as a listener for diagnostic purposes, or to simulate injection attacks on the traffic. Finally, the *Assembler* is on the right which

takes the multiple inputs and reassembles the data. The source and sink simply represent the applications which produce and receive the data that is being transmitted.

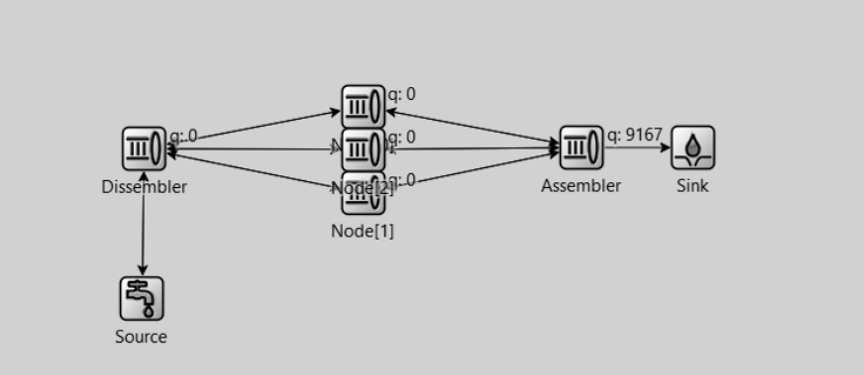


Figure 1: Network Diagram

3.3 The Dissembler

The *Dissembler* creates individual *Channel Packets* to be sent along each channel. As bits stream from the application, the *Dissembler* creates *Data Packets* which contain the unique data to be distributed across the three channels. The size of the *Data Packets* is based on the size of the *Channel Packets,* the number of channels in use and the Duplication Factor (DF) as follows:

$$\textit{Data Packet Size} = \textit{Channel Packet Size} \times (\textit{Number of Channels}/\text{DF})$$

For simplicity, a *Channel Packet* size of 64 data bits was chosen. Any empty space left by data that could not completely fill the last *Data Packet* is padded with 0's in order to maintain the *Channel* and *Data Packet* sizes. While trivially small *Data Packet* and *Channel Packet* sizes were chosen, this approach is easily scalable to larger packets, closer to the Ethernet Maximum Transmission Unit (MTU) of 1500 bytes.

3.4 Data Packets and Channel Packets

For each set of 96 bits that arrive at the *Dissembler*, populating the *Data Packets*, the *Dissembler* distributes the data into specific *Channel Packets* for each channel. *Channel Packet* A1 receives 2/3 of the bits from the *Data Packet* 1. Similarly, *Channel Packets* B1 and C1 receive 2/3 of the bits from *Data Packet* 1. Figure 2 illustrates the specific bits that go into each of the *Channel Packets*, ensuring that each bit is placed on exactly two of the three channels, thus creating a duplication factor of 2.

Data Bit	-	-	-	-	-	-	-	-	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Chan A									0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0
Chan B									0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0
Chan C									0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0
Bits A	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0
Data Index									0	2	4	3	8	7	9	11	12	14	16	15	20	19	21	23	24	26	28	27	32	31	33	35	36	38	40	39	44	43	45	47	48	50	52	51
Bits B	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	1	1	
Data Index									1	0	5	4	6	8	10	9	13	12	17	16	18	20	22	21	25	24	29	28	30	32	34	33	37	36	41	40	42	44	46	45	49	48	53	52
Bits C	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	1	0	1
Data Index									2	1	3	5	7	6	11	10	14	13	15	17	19	18	23	22	26	25	27	29	31	30	35	34	38	37	39	41	43	42	47	46	50	49	51	53
Msg Bit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43

Figure 2: Channel Packet Data Layout

Figure 2 also illustrates the structure of the *Data Packets* and *Channel Packets*. The "Data Bit" row is the index of the data from the *Data Packet*. The "Chan" rows are populated with the actual bits inside of *Data Packet* 1; note the *Data Packet* is the same across all three channels as all three channels are simply sending different parts of the same message. A particular cell highlighted in a "Chan" row indicates that bit is sent on the respective channel.

Figure 2 also illustrates the data scrambling that occurs within the protocol. This can be seen in the "Data Index" rows that correspond to each "Bits" row. The *Data Packet* bits are placed into the *Channel Packets* "Round Robin" but not sequentially. This means that the nth bit of the *Channel Packet* does not necessarily correspond to the nth bit of the *Data Packet*. Additionally, the first 8 bits are policy header bits and unrelated to data.

The final *Channel Packet* is 64 bits of headers and footers with 64 bits of data. The first 8 policy bits simply state the algorithm used by the *Dissembler* for selecting which bits were placed on which channel, as well as the scrambling code that was used. The packet number simply helps the *Assembler* ensure the correct *Channel Packets* are being assembled together. Finally, the Cyclic Redundancy Check (CRC) is a standard CRC used to check for packet validity.

	8 Bits	64 Bits	8 Bits	8 Bits	8 Bits	10 Bits	22 Bits
1	policy	data A	CS1:B[0:31]	CS2:B[32:63]	CS3:C[0:63]	packet #	CRC
2	policy	data B	CS1:C[0:31]	CS2:C[32:63]	CS3:A[0:63]	packet #	CRC
3	policy	data C	CS1:A[0:31]	CS2:A[32:63]	CS3:B[0:63]	packet #	CRC

Figure 3: Channel Packet Structure [1]

3.5 Checksums

The checksum layouts may appear peculiar at first but serve a specific purpose. The layout is based on previous work that developed this specific checksum and CRC structure so as to detect malicious injection on a channel (Simon 2022). Using this structure together with a DF of 2, data is recovered well. Figure 3 illustrates the layout of the checksums such that each channel carries a checksum for the other two channels. This allows data to be cross-checked multiple ways to ensure validity and, in the event of an error, identify invalid data and use valid data from an uncorrupted channel. This allows the *Assembler* to recover data with minimal need for retransmission.

3.6 The Assembler

The role of the *Assembler* is to take the data from each of the *Channel Packets* and reassemble the data within the *Forwarding Buffer* in order to pass the data to the application. To accomplish this, the *Assembler* must receive packets from each channel and process them. Processing the packets includes confirming packet validity, unscrambling each packet and populating the *Forwarding Buffer* with the appropriate data. If a packet is invalid, the *Assembler* attempts to recover the missing or corrupted data.

3.7 *Assembler* Operation

The *Assembler* will process the *Channel Packets* according to one of two methods. The first assumes no corruption within the channel packets and all data is valid. The second assumes corruption exists within one or more *Channel Packets.*

When the *Assembler* receives data, it assumes data is valid until proven otherwise. This means the *Assembler* waits for at least two corresponding *Channel Packets* to arrive before beginning to process the data. The *Assembler* must only wait for two of the three channels due to our duplication factor of 2, which ensures that 100% of the data is contained within just two of the channels. This assembly without all three channels is referred to as *Speculative Assembly.* When two *Channel Packets* have arrived, the *Assembler* checks the CRCs and Checksums to ensure validity of packets. Assuming data is valid, the assembler places the unscrambled data from both of the channel packets into the *Forwarding Buffer* and continues operation with the next packet in the queue. If *Speculative Assembly* was performed, the unused *Channel Packet* is dropped once it arrives at the *Assembler*.

In the event that the *Assembler* detects invalid data, assembly halts at that point until all three channels have arrived. Without all three channels, the probability of successfully recovering data is low, therefore, *Speculative Assembly* is not attempted. Once data from all three channels have arrived, the checksums are compared in order to determine the corrupt channel and utilize the data from the uncorrupted channels. In the event that no channel can be definitively identified as the corrupt channel, retransmission would be necessary. For purposes of this simulation no feedback is implemented so this need for retransmission was captured as corruption in the final message.

3.8 Metrics for Measuring Performance

The primary testing goals of this protocol are to analyse the levels of corruption in the final message to determine recovery effectiveness of this protocol as compared to a single channel system with no error correction.

Additionally, simulation execution time is analysed to assess the performance gain certain features provide the protocol, such as the speculative assembly.

3.9 The Simulation Environment

OMNeT++ is the simulation environment used to implement this protocol. Much of the infrastructure was created by Paul Simon in (Simon 2022). This includes the networking and setup of the simulation environment. All previous sections within Chapter 3 were implemented within OMNeT++ and C++.

4. Analysis

A series of experiments are executed to analyse the performance of the protocol. The results of a Single Channel systems performance, are first, followed by results for a multi-channel setup. In conclusion, we compare the two and discussion the advantages and drawbacks of each.

4.1 Single Channel Performance

Single channel performance is not a primary focus of this protocol; however, it provides a necessary benchmark. Thus, analytical results have been created to compare and benchmark the multi-channel protocol as shown in Table 1. The table shows that without any of the data recovery or redundancy that exists within the multi-channel protocol, the amount of corruption in the final message will be equal to the probability of corruption on the channel.

Table 1: Single Channel Corruption

Channel Corruption	Final Message Corruption
0	0
10	10
20	20
30	30
40	40
50	50
60	60
70	70
80	80
90	90
100	100

4.2 Multi Channel Performance

For multi-channel performance, we begin by looking at the timing performance gains provided through the use of a three-channel system with the speculative assembly feature enabled.

Table 2 shows the timings for a three-channel system with no corruption. The delay simulated for each packet along each channel is listed within the respective ``Delay" columns. The Speculative Finish and All Finish columns simply report the simulation time at which the message is received. Speculative Finish is when the program completes with speculative assembly, while All Finish is when the program completes after waiting for all packets to arrive. The difference column simply highlights the gain that speculative assembly was able to provide for the protocol.

Table 2: **Three Channel Timings**

Delay			Speculative Finish	All Finish	Difference
A	B	C			
0	0	0.001	20.953	23.388	2.435
0.001	0.001	0.001	-	23.388	-
0	0.001	0.001	-	23.388	-
0.002	0.002	0.002	-	28.002	-
0.001	0.002	0.002	-	28.002	-
0.001	0.001	0.002	23.388	28.002	4.614

Table 2 shows the performance gains achieved with speculative assembly as compared to waiting for data to arrive on all channels. This demonstrates the functionality of the protocol and suggests that as potential delays along channels increase, the amount of time saved by conducting speculative assembly greatly increases as well. This is essential because when conducting a multi-channel communication session, there is no guarantee that every channel will perform as well as the others.

Table 2 illustrates the ability for this protocol to handle complete channel failure and how speculative assembly works when the system is experiencing some level of corruption. For the example shown in Table 2, Channels A and B are each unobstructed, but Channel C is fully corrupted. No data from Channel C arrives at the *Assembler*, yet the protocol successfully reconstructs the data.

Table 3: **Three Channel System with Single Channel Failure**

Corruption			Delay			Speculative Finish	All Finish	Message Corruption
A	B	C	A	B	C			
0	0	100	0	0	0	-	20.953	0
0	0	100	0	0	0.001	20.953	23.388	0
0	0	100	0	0.001	0	-	23.388	0

These examples illustrate 1) how the protocol is effective at increasing availability of the system by remaining resilient during complete channel failures and 2) the performance gains by requiring no packet re-transmissions as a result of this channel failure.

Table 4 presents levels of corruption in the final message with corruption on all channels. With low overall channel corruption, the protocol produces final message corruption that is lower than the overall channel corruption, however when the corruption is high, the overall message corruption is higher than the channel corruption.

Table 4: **Three Channel Corruption**

Corruption			Message Corruption
A	B	C	
10	10	10	3.27
20	20	20	7.5
30	30	30	24.2
40	40	40	35.8
50	50	50	49.4
60	60	60	65.6
70	70	70	78.6
80	80	80	90
90	90	90	97.6

Figure 4 shows the corruption values from Table 1 and Table 4 plotted together. This demonstrates an interesting interaction between the two. When overall corruption levels are low on a network or channels, less than 50%, then the multi-channel system performs much better, having improvements between 1% and as much as 13%. However, when the channel corruption rises above 50%, the multi-channel system experiences a similarly worse performance over the single channel system.



Figure 4: Single-Channel vs Three-Channel Corruption

This relationship makes sense analytically, since the protocol requires at least two channels to transmit successfully. As the probability of corruption increases, the chances of at least two channels being uncorrupted decreases. However, as the probability of corruption decreases, the chance of at least two channels being uncorrupted increases. This results in the inverse relationship demonstrated.

This analysis demonstrates areas and situations where a multi-channel setup is superior to a single channel setup. When corruption levels are greater than 0 but less than 50%, the multi-channel setup is superior to a single channel setup with regards to the success of data transmission. In a realistic scenario where feedback was implemented, this would also result in fewer re-transmissions required. While it was shown that, without corruption, the single channel system is superior to the multi-channel from a bandwidth perspective. Once corruption is involved and the single channel system would need to retransmit 5% to 10% more packets than the multi-channel, the multi-channels' benefits emerge.

Additionally, the additional security benefits of data scrambling and dropout come at no cost when implementing a protocol across three channels with a duplication factor of two. Should an adversary attempt to eavesdrop on the transmissions of a single channel, not only will the bit level scrambling make the data unreadable without the scrambling code, but the missing data will make determining this scrambling code significantly more difficult. In order to reasonably determine the whole message; an adversary would need to sniff on at least two of the channels in order to have all of the data to decode. However, the added problem of too much information also complicates the problem. Sorting out the “chaff” from the unique data would require significant time and effort, and this comes at no additional cost to the user, aside from the standard multi-channel overhead.

Finally, the overall availability of the multi-channel system is clearly superior to a single channel system. With a single channel, if the channel fails, then the entire system will not work. A channel may fail for many reasons, from something benign like maintenance on a channel or a cable being accidentally cut, to something malicious like a Denial of Service or Distributed Denial of Service attack. Regardless of the cause of an outage, a single channel system would be unable to operate until the cause of the outage has been resolved, or some alternative has been worked out. With the multi-channel protocol, a failure of a single channel for any reason is completely abstracted away from the user and application. They would be completely unaware of the issues faced by the network because there would be, at most, a slight increase in the delay the network traffic experiences; as demonstrated through Table 2.

4.3 Results Summary

Overall, it has been shown that a multi-channel protocol as described in this research is able to produce lower overall corruption in a contested environment, assuming corruption levels are below 50%. Additionally, the benefits of speculative assembly have been shown within the context of the protocol described. While some drawbacks to the protocol exist, when operating within a contested environment, the security and performance benefits of the protocol illustrate the benefits of using a multi-channel system.

5. Conclusion

While there are numerous benefits to adopting a multi-channel strategy such as the one described within this research, there are also numerous reasons to avoid such a strategy. This section examines what was shown within this research, the significance of the findings, and suggests opportunities for future work.

5.1 Research Conclusions and Significance

The primary contribution of this research is demonstration of the benefits of a multi-channel protocol when operating in a contested environment at less than 50% channel corruption. Under these conditions, the data corruption experienced by the final message is anywhere from 5% to 13% less than a comparable single channel system.

Additionally, the security benefits provided by a multi-channel protocol that implements data dropout and scrambling is clear. The challenges to an eavesdropper attempting to read the original message are significant due to the removal of information available to the attacker. Together with scrambling, this data obfuscation results in significantly increased security over a single channel system with lower processing and overhead than encryption.

The development of this simulation environment and protocol creates the opportunity for further research into multi-channel protocols and further exercising this particular protocol to explore additional benefits or features.

5.2 Future Work

The primary area for future work is advancing features of the protocol within the simulation environment, as well as to more thoroughly exercise the simulation in order to better characterize the performance and abilities of the protocol.

The implementation of an initialization vector or handshake would be valuable. The way to define and coordinate channels would be essential for smooth operation of the protocol. Additionally, in order to maintain the security of the protocol, this initialization must happen in a discreet way so that adversaries are unable to gain information regarding where additional channels are, what protocols are being used or any other information that could be beneficial towards the deciphering of the messages. One potential solution would be to utilize PKI to encrypt the initial handshake to ensure it remains secret, then use the protocol as described for the remaining data transfer, so as to not incur the overhead of encryption with the remaining data transfer.

Exercising the simulation more thoroughly would also be a beneficial exercise to generate more data on the performance of the protocol in specific edge cases or with larger data sets. Combining this with a feedback and re-transmission capability would facilitate a realistic comparison to a single channel system.

5.3 Conclusion

While no perfect solution can ever exist in a complex and ever-changing problem, this research seeks to create a viable solution for a specific scenario by utilizing multi-channel communications. Through the use of data obfuscation via bit scrambling and data dropout, the challenges of a potential attacker have been made significantly greater. Additionally, performance gains in non-permissive environments have been demonstrated, showing how a multi-channel protocol is superior to a single channel when data corruption rates are below 50%. Finally, practical applications and areas for future focus and work have been discussed to illustrate the significance of this research and the way for it to move forward. Future work could include the progression of the simulation to better encompass a robust and complete protocol, or implementation within a network that has the opportunity to utilize multiple channels.

References

Bao, S.-D., Chen, M., & Yang, G.-Z. (2017). A Method of Signal Scrambling to Secure Data Storage for Healthcare Applications. IEEE Journal of Biomedical and Health Informatics, 21(6), 1487–1494. https://doi.org/10.1109/JBHI.2017.2679979

Bonaventure, O., & Seo, S. (2016). Multipath TCP Deployments. IETF.

Dmitrienko, A., Liebchen, C., Rossow, C., & Sadeghi, A.-R. (2014). On the (In)Security of Mobile Two-Factor Authentication. In N. Christin & R. Safavi-Naini (Eds.), Financial Cryptography and Data Security (pp. 365–383). Springer Berlin Heidelberg.

Ford, A., Raiciu, C., Handley, M., Bonaventure, O., & Paasch, C. (2020). TCP Extensions for Multipath Operation with Multiple Addresses (RFC No. 8684; pp. 1–68). RFC Editor. https://www.rfc-editor.org/rfc/rfc8684.txt

Hughes, G. F., & Murray, J. F. (2005). Reliability and Security of RAID Storage Systems and D2D Archives Using SATA Disk Drives. ACM Trans. Storage, 1(1), 95–107. https://doi.org/10.1145/1044956.1044961

Modarressi, A. R., & Skoog, R. A. (1990). Signaling System No.7: a tutorial. IEEE Communications Magazine, 28(7), 19–20. https://doi.org/10.1109/35.56239

Simon, P. (2022). Development of a Security-Focused Multi-Channel Communication Protocol and Associated Quality of Secure Service (QoSS) Metrics [Phdthesis]. Air Force Institute of Technology.

Simon, P. M., Graham, S., Talbot, C., & Hayden, M. (2021). Model for Quantifying the Quality of Secure Service. Journal of Cybersecurity and Privacy, 1(2), 289–301.

What is OMNeT++? (2022). https://omnetpp.org/intro/

UAV Payload Identification with Acoustic Emissions and Cell Phones

Hunter G. Doster, Barry E. Mullins Ph.D. P.E.
Air Force Institute of Technology, Wright-Patterson AFB, Ohio

hunter.doster.1@us.af.mil
Barry.Mullins@afit.edu

Abstract: The growing presence of Unmanned Aerial Vehicles (UAVs) in all sectors of society poses new security threats to civilian and military sectors. In response, new UAV detection systems have and are being developed. Current systems use techniques such as Radio Detection And Ranging (RADAR), visual recognition, and Radio Frequency (RF). Another promising solution for UAV detection uses acoustic emissions. Past researchers demonstrated the ability to use UAV acoustic signatures to determine whether a UAV carries a payload and the weight of that payload at close range with high-quality microphones. This research expands the field of study by performing acoustic payload detection using cell phones and at farther range by developing the system called HurtzHunter. The system collects audio data and extracts Mel-Frequency Cepstrum Coefficients (MFCCs) to train Support Vector Machines (SVMs). The HurtzHunter system tests acoustic payload detection with one high-quality microphone and six different cell phones at 7 m - 100 m ground distance from the UAV. At each distance, the experiment runs 6 flights each with a unique payload attached to the UAV. The HurtzHunter design achieves an 88.26% - 99.93% payload prediction accuracy depending on the configuration.

Keywords: Cell Phone, UAV Detection, Mel-Frequency Cepstrum Coefficients, Support Vector Machine, Acoustic Emissions

1. Introduction

Today, there are over 1.25 million Unmanned Aerial Vehicles (UAVs) in operation ranging from $30 to over $35,000 for recreation and industry purposes (FAA 2019) (Kolamunna 2021). This technology brings much excitement, but as with any new technology, UAVs bring new security threats. In 2017, Syrian forces found an ISIS UAV Improvised Explosive Device (IED) plant. The terrorist organization used captured US UAVs to develop their own UAVs and turn them into IEDs (Snow 2017). Moreover, illegal drug traders use small UAVs to ship drugs across the US southern boarder (Wright). In response to these threats, there is an increasing need for UAV detection systems.

1.1 Problem Statement

Current UAV detection platforms focus on Radio Detection And Ranging (RADAR), image recognition, Light Detection And Ranging (LiDAR), or Radio Frequency. Although these platforms are useful, they can be spoofed, and some are limited to direct line-of-sight. Therefore, the need for a new zero-touch, zero line-of-sight acoustic detection system would be useful since UAVs emit acoustics which are more difficult to spoof, do not require line-of-sight, and can be fingerprinted. An effective acoustic payload detection system would allow the identification of a hazardous UAV when other methods do not since heavier payloads stress the UAV's motors more.

1.2 Past Research

Previous work demonstrates the ability to detect the presence of a UAV, the position of a UAV, as well as the payload carried by a UAV using acoustic emissions (Anwar 2019) (Bernardini 2017) (Jeon 2017) (Kim 2018) (Yue 2018). In Traboulisi et al (2020), researchers developed a UAV payload identification system using MFCCs and a Long Short-Term Memory (LSTM) recurrent neural network. For all experiments, the Parrot Mambo Mini Copter was flown indoors with payload weights from 0 grams (g) to 28 g attached. Acoustic emissions were recorded indoors at various ranges (exact distance is not specified in study) using the microphone built into a late 2013 Model Macbook pro. Each sample was processed in MATLAB using a package with a built in LSTM model. Throughout experimentation, white noise from the Gaussian model was added to test different SNR ratios. Their methodology achieved an 87.5% payload prediction accuracy.

In Ibrahim et al (2020), researchers used MFCC components with various Machine Learning (ML) algorithms for UAV acoustic payload identification. Their system collected acoustic emissions with the RodeVideoMicPro (RVMP) at 7 m ground distance from the 3DR Solo UAV. Of the 10 ML algorithms tested for acoustic payload

prediction, the Support Vector Machine (SVM) performed the best, achieving an average payload classification accuracy of 98.9%.

At the time of this research, we have not identified any other studies which explore the ability of cell phones to classify UAV payloads from acoustic emissions. This research demonstrates that cell phones can detect and classify UAV payloads from acoustic emissions. In addition, we have not identified any other studies which explore the maximum range for acoustic payload detection. This research tests acoustic payload detection from 7 m to 100 m ground distance, providing novel insight into the maximum range for acoustic payload identification.

2. Technology

This research extracts MFCCs from recorded UAV acoustics, uses the MFCCs to train SVMs for payload identification, and tests the SVMs payload prediction accuracy.

2.1 Mel-Frequency Cepstrum Coefficients

In 1980, Davis and Mermelstein designed MFCCs to digitally describe how the human phonetic system perceives an audio signal (Davis 1980). Today, systems use MFCCs for wireless audio sensor networks, multi-media retrieval, detecting biodiversity metrics in natural environments, terrorist threat detection, home invasion, and much more (Bonet-Sola 2021). MFCCs are a series of coefficients whose values give specific information of what is present in a sound. For example, Artificial Intelligence (AI) uses MFCCs to digitally perceive what a person is saying. Figure 1 shows the process for creating MFCCs.

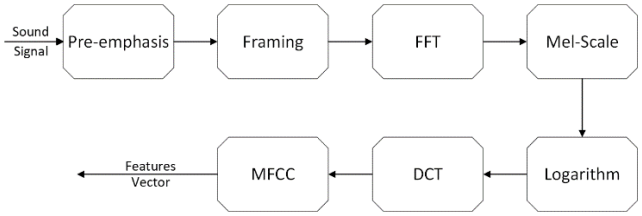


Figure 1: MFCC Generation Process (Ibrahim 2020)

First, acoustic emissions are collected by a recording device; this audio is converted to a digital form. Figure 1 calls this step Pre-emphasis because different microphones prepare the acoustic emissions differently. Once in digital form, the acoustic data is divided into sound constant frames. The frequencies in a sound constant frame do not change. Then a Fast Fourier Transform (FFT) places each frame on a spectrum. After, the spectrums are mel-scaled to reflect human-perceptual changes in frequency rather than just any change in frequency. Next, the logarithm of the amplitude is taken to convert the loudness of each spectrum in a human-perceptually relevant way. Lastly, the Discrete Cosine Transform (DCT) extracts MFCCs from each frame.

2.2 Support Vector Machine

An SVM is a type of ML algorithm which uses hyperplanes to separate classes of data (Kowalczyk 2017). The goal of an SVM is to fit hyperplanes to the data which maximize the margin between classes. Figure 2 shows an SVM with a linear hyperplane.

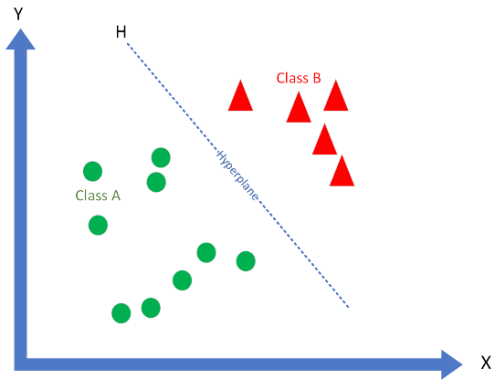


Figure 2: Basic Linear Support Vector Machine

In Figure 2, there is data from two different types of classes. The green circles are data of one class (Class A), and the red triangles are data of another class (Class B). The dotted line is the hyperplane. Each data point has a set of features (x and y) which places it on one side of the line; based on the side it falls on, the SVM knows to which class it belongs. Figure 3 shows the components an SVM uses to generate hyperplanes and learn the difference between classes.

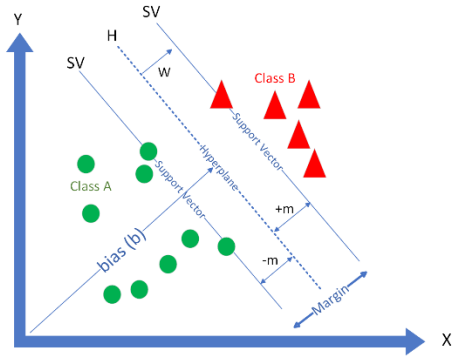


Figure 3: Detailed SVM Graph

The Support Vectors (SVs) are the lines determined by the closest data points from each class to the Hyperplane (H). W (weight) is the normal vector to the hyperplane. The distance from the SVs to the hyperplane is -m and +m. The total distance between the classes is the margin, and the bias (b) is the distance from the hyperplane to the origin. When creating an SVM, the user introduces training data to an SVM with its class identity. In the case of Figure 3, the user introduces the green circles with the identity Class A and introduces the red triangles with the identity Class B. Using the features of each data point (x and y), the SVM plots the data in 2-D space. From here, the SVM fits a linear hyperplane and SVs between the classes which maximize the margin between the classes. After determining the hyperplane and SVs, the SVs satisfy

$$w * x + b = -1 : \text{For Class A (1)}$$
$$w * x + b = 1 : \text{For Class B (2)}$$

where x is a vector of the classifiers x and y used to plot the data point, w is the weight, and b is the bias (Kowalczyk 2017). With SVs established, the SVM can categorize new data points using

$$w * x + b = output \text{ (3)}$$

where output determines classification: Class A if less than or equal to -1 and Class B if greater than or equal to 1.

Figures 2 and 3 show data which perfectly fits a hard-margin linear SVM. Hard-margin means there are no outliers, and all data perfectly fits on the correct side of the hyperplane. Due to noise, most experimental data will not fit a hard-margin SVM, so a soft-margin SVM is used. In a soft-margin SVM, the ML is allowed to make mistakes to accommodate outlier conditions which do not fall on the proper side of the SVs. A penalty is given for each mistake, and the SVM chooses a hyperplane which minimizes the number of penalties while maximizing the margin between the data. Even with soft-margin SVMs, a linear hyperplane is often not sufficient to fully classify data, so the kernel trick is commonly employed. Instead of using a linear hyperplane, the kernel trick allows the SVM to use hyperplanes with greater dimensions. Figure 4 shows a linear kernel SVM, and Figure 5 shows a sixth degree polynomial kernel SVM. In Figures 4 and 5, the two different classes are the blue stars and the red triangles. The linear hyperplane was not adequate to properly separate the two classes–a sixth degree polynomial hyperplane is required. Hyperplanes with higher dimensonality are required because many classifiers do not follow linear behavior. SVM kernels exist in every type of polynomial or dimensions. Although the kernel trick fixes this problem, hyperplanes with greater dimensonality create a greater risk of overfitting the data.

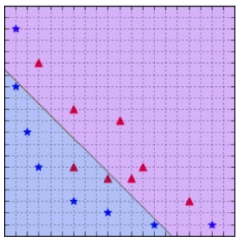


Figure 4: Linear Kernel SVM (Kowalczyk 2017)

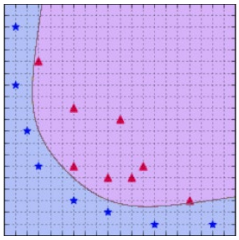


Figure 5: Sixth Degree Polynomial Kernel SVM (Kowalczyk 2017)

3. Methods

This research extends the field of study by focusing on the following questions:

- Can cell phones detect UAV payload weight with sound?
- What is the maximum range of acoustic UAV payload detection?
- Does a high-quality microphone perform better than a cell phone for acoustic payload detection?

3.1 HurtzHunter Design

As shown is Figure 6, HurtzHunter is a UAV detection system which characterizes UAV payloads. Specifically, it uses the acoustic emissions from UAVs to determine the UAV's payload weight. As shown in Figure 7, HurtzHunter conducts payload classification at varied distances and with cell phones. It records acoustic stimulus at 7 m - 100 m from the UAVs. The system uses the following recording devices: three Samsung Galaxy S8s (GS8), one Samsung Galaxy S20 (GS20), one Google Pixel 4 (GP4), one iPhone 11 Pro (iP11P), one RodeVideoMicPro (RVMP). The researchers in (Ibrahim 2020) successfully used the RVMP to create a UAV payload prediction system for close distances; therefore, this research incorporates the RVMP to provide experiment validation and comparison with past research. Unlike the cell phones, the RVMP cannot record and store audio without external audio software and storage. So, the microphone connects to an HP Pavilion business laptop with a 10th Gen Intel Core i5 during recording, and the design stores the audio directly onto the laptop using Audacity (HP) (Crook). To connect the microphone to the laptop, the design uses the TechRise USB External Stereo Sound Adapter Splitter Converter as an analog-to-digital converter (Newegg). The cell phones do not need to be connected to anything while recording.

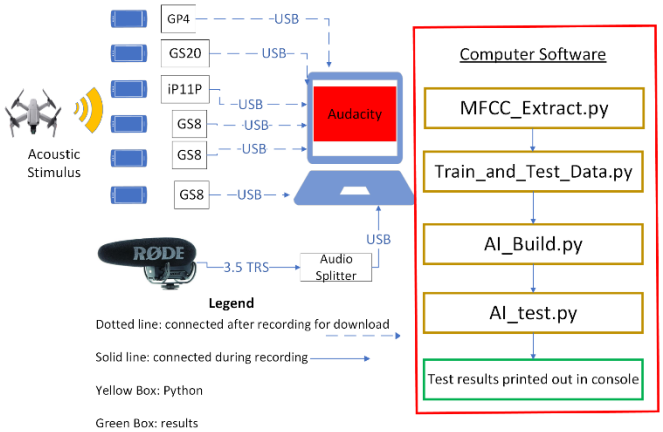


Figure 6: HurtzHunter Design (DJI)

The first phase is recording. The only device connected to the laptop during recording is the RVMP. After recording, the system enters phase two where the system stores the audio recordings on the laptop. In phase two, the user manually connects each cell phone to the laptop one-by-one and downloads the audio files to the laptop. Then, the user manually extracts the RVMP recordings from Audacity to the laptop's file system. Once all the audio files from each device are in the laptop's file system, phase 3 begins. In phase 3, Python version 3.8.8 scripts build SVMs for each recording device from their audio recordings (Python). The user manually runs the scripts in JupyterNotebook (Jupyter Notebook). MFCC_Extract.py uses the Python library librosa to extract the MFCCs for each audio recording and stores the MFCCs in a Pickle file. Then, Train_and_Test_Data.py imports the MFCCs, divides the MFCCs into training and testing data using 5-cross validation, and stores the training and testing data in a Pickle file. AI_Build.py uses the training data to build and train SVMs for acoustic payload prediction using the Python library sklearn. Once trained, the script stores the SVMs to Pickle files. AI_Test.py tests the SVMs with the testing data and prints the prediction results to the console. HurtzHunter uses a 40 millisecond (ms) frame length and 20 ms hop length. This research records acoustic emissions for 60 seconds (s). Therefore, the number of frames for each audio recording is 2999 derived by

$$\#\ of\ frames = \frac{recording\ time}{hop\ length} - 1 = \frac{60}{0.02} - 1 = 2999 \quad (4)$$

In using 5-cross validation, 80% of the data is for training, and 20% is for testing. Each frame results in 1 array of MFCCs resulting in 2999/5 = 599.8 (floored to 599) full MFCC arrays used for testing. Therefore, prediction accuracy (AI accuracy) is (# of correct predictions) / 599.

3.2 Experiments

Figure 7 illustrates the experiment.

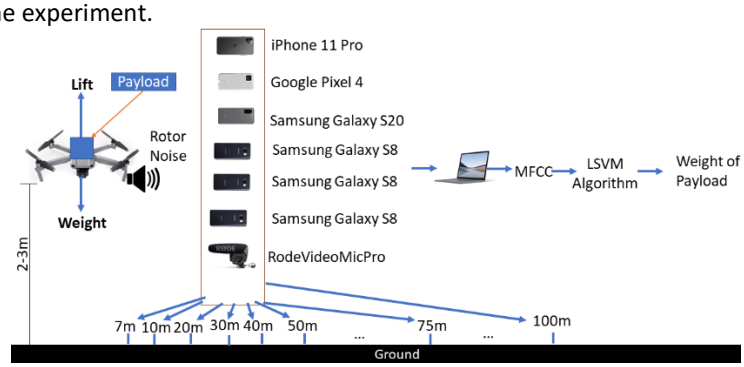


Figure 7: Experiment Diagram (DJI)

All data was collected on a flight line at the Air Force Bombing Range in Avon Park, FL. In the experiment, a DJI Mavic Air 2 hovers 2 m - 3 m above the ground (DJI). While hovering, the 7 microphones simultaneously collect acoustic emissions for 60 s. There are 48 flights in total from which the microphones collect data. The flights are conducted at 7 m, 10 m, 20 m, 30 m, 40 m, 50 m, 75 m, and 100 m ground distance from the microphones. At each distance 6 flights are flown each with a different payload attached to the top of the UAV. The payload weights are 0 g, 56 g, 112 g, 168 g, 224 g, 280 g. The raw data from this experiment is used in 3 different case studies each providing answers to the research questions listed at the beginning of Section 3 Methods. During experimentation, background noise from birds, passing cars, bugs, and wind were present. Wind noise is a major nuisance factor which affected the results discussed in Section 5 Nuisance Factors.

4. Results and Discussion

This academic paper uses the data collected by the RVMP and one of the GS8s. Future work will include the data from the other recording devices shown in Figure 7.

4.1 Case Study 1

The purpose of this case study is to provide experiment validation by repeating experiments in past research Ibrahim et al (2020), and expand the research field by using a cell phone as the recording device. The experiments in Ibrahim et al (2020) used the RVMP as the recording device and collected acoustic emissions from a hovering UAV at 7 m ground distance from the UAV. This study repeats the experiment with the RVMP and one GS8. This case study provides evidence to answer the following questions: "Does a high-quality

microphone perform better than a cell phone for acoustic payload detection?" and, "Can cell phones detect UAV payload weight with sound?". This case study uses three SVMs, one with a linear, a quadratic, and a cubic kernel because the experiments in Ibrahim et al (2020) used these SVM kernels. In Ibrahim et al (2020), the researchers achieved prediction accuracy ranging from 98.5 - 98.9 % with their configurations. Figure 8 shows the results from this case study.

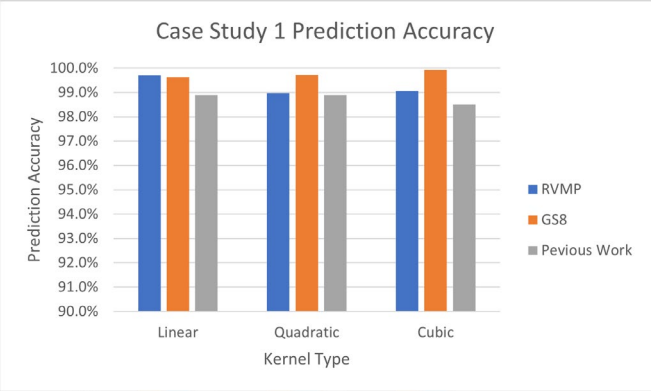


Figure 8: Case Study 1: Prediction Accuracy Results And Comparison

Using the Linear SVM (LSVM), the GS8 achieves 99.62% total accuracy, the RVMP achieves 99.71% average accuracy, and the researchers in Ibrahim et al (2020) achieved 98.9% accuracy. The accuracy for the GS8 and RVMP is the average accuracy across all payload weights. Using the Quadratic SVM (QSVM), the GS8 achieves 99.72%, the RVMP achieves 98.98% accuracy, and the researchers in Ibrahim et al (2020) achieved 98.9% accuracy. Using the Cubic SVM (CSVM), the GS8 achieves 99.93%, the RVMP achieves 99.06% accuracy, and the researchers in Ibrahim et al (2020) achieved 98.5% accuracy. Both the RVMP and GS8 consistently perform with nearly the same accuracy as past research. These results indicate the experiment behaves consistently with past research providing evidence for experiment validation. In addition, the results show that the GS8 performs at the same level as the RVMP. So, there is evidence to suggest a cell phone device achieves the same performance as a high-quality microphone for acoustic payload detection at 7 m. Lastly, the three SVM configurations (linear, quadratic, and cubic kernel) perform nearly the same for the GS8, RVMP, and prior research (Ibrahim 2020).

Table 1: Case Study 1 Runtimes

		1 AI (s)	5 AI (s)
RVMP	LSVM	0.14	0.98
	QSVM	0.52	2.92
	CSVM	0.4	2.43
GS8	LSVM	0.13	0.72
	QSVM	0.38	2.54
	CSVM	0.31	1.69

Table 1 shows the runtime of 1 AI and 5 AIs for each configuration. The reason this study shows 5 AIs is because 5-cross validation requires 5 AIs for implementation. So, the runtime to build all 5 AIs conveys the real runtime to build the AI for HurtzHunter. The runtime for the LSVM is shorter than the QSVM and CSVM for both the GS8 and RVMP. However, the longest runtime is 2.4 s which is relatively short, and does not affect the goals of this research.

4.2 Case Study 2

Case Study 2 builds upon Case Study 1 by increasing the distance between the microphone and UAV from 7 m to 100 m. Previous work gathered acoustic emissions at 7 m away from the UAV. This research expands the field of study by gathering acoustic emissions at 7 m to 100 m away. This study uses the one GS8 (from Case Study 1) and the RVMP to collect acoustic data at 7 m, 10 m, 20 m, 30 m, 40 m, 50 m, 75 m, and 100 m ground distance from the UAV. Additionally this case study builds an LSVM, QSVM, and CSVM for each recording device. The

results in this case study provide answers to the questions "What is the maximum range of acoustic UAV payload detection?" and, "Does a high-quality microphone perform better than a cell phone for acoustic payload detection?" Figure 9 shows the prediction accuracy for all three SVMs using the RVMP.

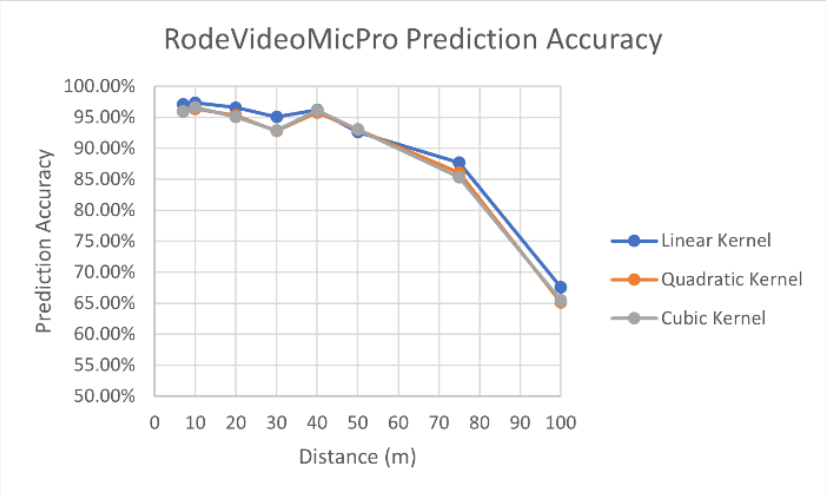


Figure 9: Case Study 2: RVMP Three SVM Comparisons

Using the RVMP, the average accuracy across all distances and payloads for the LSVM is 91.27%, for the QSVM is 90.05%, and for the CSVM is 90.05%. As Figure 9 shows, the prediction accuracy remains above 90% from 7 m - 50 m, then declines. Case Study 1 shows a total accuracy of 99%. When all distances are included, the average accuracy decreases to 90.05% - 91.27%. This decrease is expected due to the loss of signal strength as distance increases. A sound is fainter at longer distances and stronger at shorter distances, and fainter sounds are more difficult to capture with a microphone. Therefore, a weaker signal should lead to a decrease in prediction accuracy because it carries less information. The results in Figure 9 support this phenomenon as the percent accuracy decreases as distance increases. Specifically, the main decrease is from 50 m to 100 m. In addition, the signal strength does not affect the prediction accuracy until after 50 m. Therefore, there is evidence to suggest the maximum range for acoustic payload detection is 50 m.

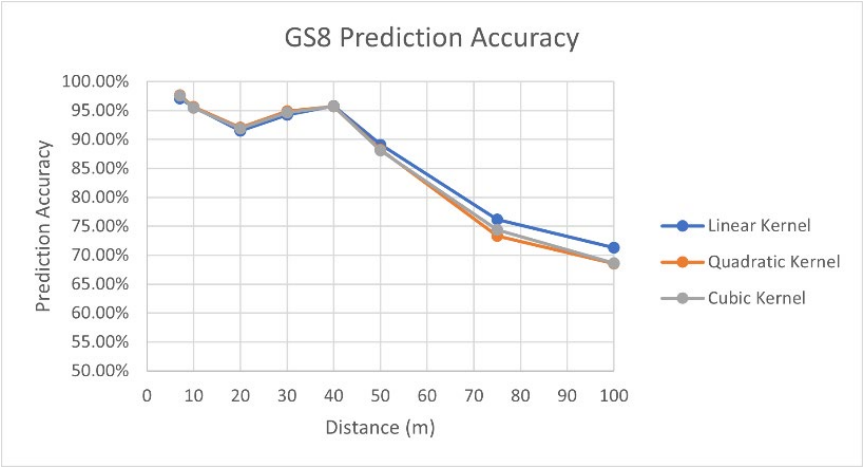


Figure 10: Case Study 2: GS8 Three SVM Comparison

Figure 10 shows the prediction accuracy for all three SVMS using the GS8. The average accuracy across all distances and payloads for the LSVM is 88.84%, for the QSVM is 88.26%, and for the CSVM is 88.30%. As Figure 10 shows, the prediction accuracy remains above 90% from 7 m - 40 m, then declines. Case Study 1 shows a total accuracy of 99%. When all distances are included, the average accuracy decreases to 88.26% - 88.84%. As observed from the GS8, loss of signal strength leads to a decrease in prediction accuracy, but there is not a significant drop in accuracy until after 40 m.

Table 2 shows the runtimes for the RVMP, and Table 3 shows the runtimes for the GS8.

Table 2: Case Study 2 RVMP Runtimes

		RVMP Runtime		
		Seconds	Minutes	Hours
Linear	1 AI	3781.8	63	1.1
	5 AI's	19971.6	332.9	5.5
Quadratic	1 AI	224.6	3.7	0.1
	5 AI's	1086.5	18.1	0.3
Cubic	1 AI	205.6	3.4	0.1
	5 AI's	1061.6	17.7	0.3

Table 3: Case Study 2 GS8 Runtimes

		GS8 Runtime		
		Seconds	Minutes	Hours
Linear	1 AI	7831.9	130.5	2.2
	5 AI's	34541.9	575.7	9.6
Quadratic	1 AI	257.6	4.3	0.1
	5 AI's	1357.5	22.6	0.4
Cubic	1 AI	264.5	4.4	0.1
	5 AI's	1324.1	22.1	0.4

In Case Study 1, the runtime was less than 3 s. In this Case Study, the runtime to fully classify a LSVM increases to 5.5 hours (hrs) for the RVMP and 9.6 hrs for the GS8. The reason for this change is because the AI in this study uses data from all ranges (7 m to 100 m) which is seven times the data Case Study 1 uses. In addition to the amount of data, the nature of the data in this case study is more difficult to classify than the data in Case Study 1. Due to factors such as range, loudness, and wind noise, the SVM needs more hyperplanes to fully classify the data. This result indicates a higher degree kernel is ideal to fully classify the SVM.

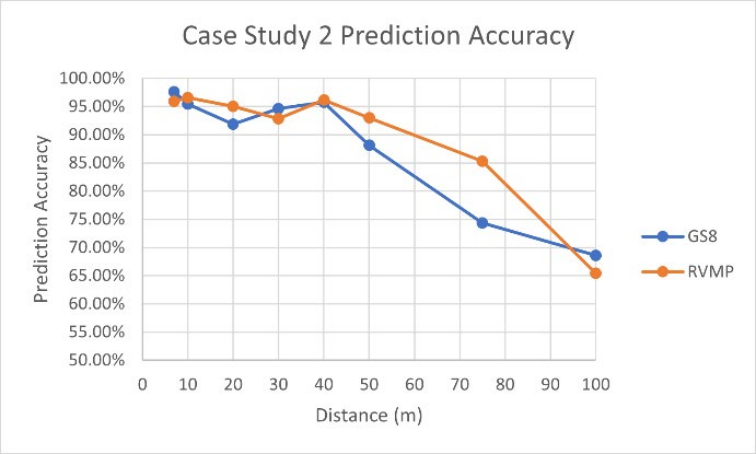


Figure 11: Case Study CSVM Comparison

The Quadratic and Cubic SVMs have higher dimensionality, so their hyperplanes are able to better classify the data. Because all three kernel configurations for the SVM follows the same results, this research concludes the CSVM is the optimal configuration due to its high accuracy and short runtime. Therefore, this section compares the CSVM for the RVMP and GS8. Figure 11 shows the CSVM for the RVMP and GS8. The graph shows that there is not an optimal device for all distances. It shows that the cell phone performs relatively the same as a high-quality microphone. Additionally, both devices observe a drop-off in prediction quality after 40 m.

5. Nuisance Factors

The three largest nuisance factors are wind, time of day, and battery life. Wind and time of day affect the prediction results. These factors were uncontrollable in the experiment. Only two UAV batteries were available and one battery charger available on the flight line. Each battery provides ~12 min of flight time, and it takes 2 hours to charge a battery. So, the recordings were spread out over the day. The different times of flight affected the experiment results because the winds varied throughout the day. Recordings done during high winds show lower prediction accuracy than those during low winds. Figure 12 shows the prediction accuracy over time of day. The figure shows results from the GS8 in Case Study 2. All devices show this same trend.

The order of weight classes was randomized. The order is as follows: 280 g, 0 g, 168 g, 224 g, 56 g, 112 g. Therefore 280 g and 0 g were ran in the morning, 224 g was ran around noon, 56 g was ran in the early afternoon, and 112 g was ran in the late afternoon. Winds were calm in the early morning and began to pick up around mid morning peaking in the early afternoon; they started to trail off in the mid afternoon. Figure 12 shows how accuracy changed throughout the day and that there is a direct correlation between wind and accuracy. Table 4 shows the wind noise readings for each weight. The experiments use the iOS National Institute for Occupational Safety and Health (NIOSH) Sound Meter app to measure ambient noise (EA Lab). The Center for Disease Control (CDC) developed this app to measure ambient noise in dB (CDC). Right before each flight, the researcher used this app on the iP11P to measure the ambient noise in dB.

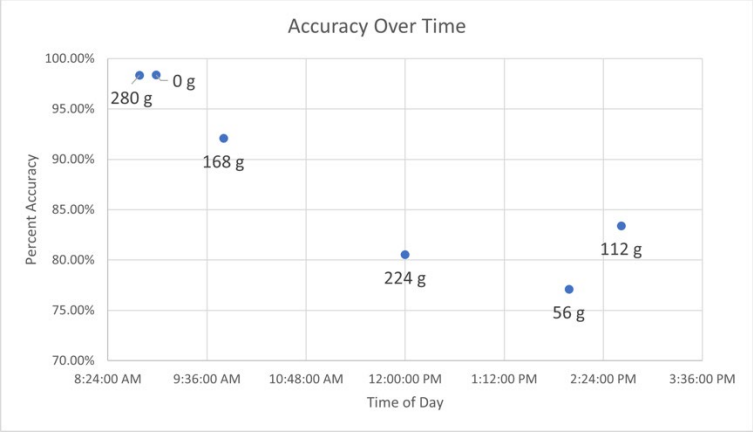


Figure 12: Prediction Accuracy Over Time of Day

The results shows wind noise has a clear effect on the system. When the noise level is below 50 dB, the system remains fairly accurate, but at or above 54 dB there is a significant decrease in accuracy. The wind noise begins todrown out the UAV acoustic emissions somewhere between 50 dB and 54 dB. Further research needs to find the exact threshold.

Table 4: Wind Noise

Weight (g)	Wind Noise (dB)	Time of Day
280	44	8:47:00 AM
0	45	8:59:00 AM
168	49.7	9:48:00 AM
224	55.4	12:00:00 PM
56	62.5	1:59:00 PM
112	54.7	2:37:00 PM

6. Conclusion

This research demonstrates the ability to use UAV acoustic emissions to identify payload weight at up to 100 m away from the UAV with cell phones. This research developed the HurtzHunter design to demonstrate the ability to use cell phones for payload detection. Based on the configuration, the procedures demonstrated in this research are capable of achieving a payload prediction accuracy between 88.26 - 99.93% depending on the configuration.

Acknowledgments

This Cyber Physical Sensing research was sponsored by the Air Force Research Laboratory, Sensors Directorate, Sensing Effects & Analysis Branch (RYAA). Cyber Physical Sensing is defined as sensing the induced effect from the physical world through the cyber-domain. The effect only exists as a product of cyber-systems and their interaction.

References

Federal Aviation Administration (2019) “Faa Aerospace Forecast 2019- 2039” Federal Aviation Administration, [Online], https://www.faa.gov/data research/aviation/aerospace forecasts/media/FY2019-39 FAA Aerospace Forecast.pdf Accessed Jan. 4, 2022

(2019) “Navigating Drone Prices: A Guide to What to Expect at Different Price Points When Buying a Drone”, [Online], https://uavcoach.com/drone-prices/ Accessed: Jun. 3, 2021

H Kolamunna et al, (2021) “DronePrint: Acoustic Signatures for Openset Drone Detection and Identification with Online Data” Proceedings of the ACM on Interactive, Movile, Wearable and Ubiquitous Technologies, vol. 5, no. 1, pp. 1-31, [Online], https://dl.acm.org/doi/abs/10.1145/3448115/ Accessed Jan. 21, 2022

S. Snow (2017) “Syrian fighters in Raqqa uncover ISIS drone factory”, [online], https://www.militarytimes.com/flashpoints/2017/07/26/syrian-fightersin-raqqa-uncover-isis-drone-factory/ Accessed: Dec. 10, 2021

T. Wright , “How many drones are smuggling drugs across the US southern border?” [Online], https://www.airspacemag.com/flighttoday/narcodrones-180974934/ Accessed: Dec. 10, 2021

M. Anwar, z. Kaleem, A. Jamalipour (2019) “Machine Learning Inspired Sound-Based Amateur Drone Detection For Public Safety Applications,” IEEE Transactions on Vehicular Technology, Vol. 68, no. 3, pp. 2526-2534, [online] https://ieeexplore.ieee.org/abstract/document/8616877 Accessed: Jan. 21, 2022

A. Bernardini, F. Mangiatordi, E. Pallotti, L. Capodiferro. (2017) “Drone detection by acoustic signature identification,” IS and T International Symposium on Electronic Imagining Science and Technology, pp. 60-64 [online] https://www.ingentaconnect.com/content/ist/ei/2017/00002017/00000010/art00009 Accessed: Jan. 21, 2022

S. Jeon, J. Shin, Y. Lee, W. Kim, Y. Hyoun, H. Yang (2017) “Empirical study of drone sound detection in real-life environment with deep neural networks,” 25th European Signal Processing Conference, EUSIPCO pp. 185-1862, [online], https://ieeexplore.ieee.org/abstract/document/8081531 Accessed: Jan. 21, 2022

J. Kim, D. Kim (2018) “Neural Network based Real-time UAV Detection and Analysis by Sound,” Journal of Advanced Information Technology and Convergence, Vol. 8, No. 1, pp. 43-52. [Online] https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE07504681 Accessed: Jan. 21, 2022

X. Yue, Y. Lui, J. Wang, H. Song, J. Cao (2018) “Software Defined Radio and Wireless Acoustic Networking and Amateur Drone Surveillance,” IEEE Communications Magazine Vol. 56, No. 4, pp. 90-97. [Online], https://ieeexplore.ieee.org/abstract/document/8337902 Accessed: Jan. 21, 2022

A. Traboulisi and M. Barbeau (2020) “Identification of Drone Payload Using Mel-Frequency Cepstral Coefficients and LSTM Neural Network,” Proceedings of the future technologies conference Vol. 1, pp. 402-412. [Online] https://link.springer.com/chapter/10.1007/978-3-030-63128-4 _30 Accessed: Jan. 21, 2022

O. Ibrahim, S. Sciancalepore, and R. Pietro (2020) “Noise2Weight: On Detecting Payload Weight from Drone Acoustic Emissions,” arXiv Audio and Speech Processing, pp. 1-14. [Online], https://arxiv.org/abs/2005.01347 Accessed: Jan. 14, 2022

S. B. Davis and P. Mermelstein (1980) “Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences” Transactions on Acoustics, Speech, and Signal Processing Vol. 28, No. 4, pp. 357-366. [Online] https://ieeexplore.ieee.org/abstract/document/1163420 Accessed: Jan. 21, 2022

D. Bonet-Sola and R. M. Alsina-Pages (2021) “comparative survey of feature extraction and machine learning methods in diverse acoustic environments” Sensors (Switzerland), Vol. 21, No. 4, pp. 1-21. [Online] https://www.mdpi.com/1424-8220/21/4/1274 Accessed: Jan. 21, 2022

A. Kowalczyk (2017), Support Vector machines, Syncfusion, North Carolina, [E-books]. https://www.syncfusion.com/succinctly-freeebooks/support-vector-machines-succinctly Accessed: Jan. 14, 2022

J. Crook (2021) “Audacity Home Page,” Audacity [Online] https://www.audacityteam.org Accessed: Dec. 16, 2021

Newegg, “USB Sound Card, TechRise USB External Stereo Sound Adapter Splitter Converter” newegg, [online] https://www.newegg.com/p/1B4-08DC-001A5 Accessed: Dec. 16, 2021

HP, “HP Pavilion 15”, HP [Online], https://www.hp.com/usen/shop/mdp/laptops/pavilion-15-344522‑#!&tab=vao Accessed: Jan. 21, 2022

DJI, “Mavic Air 2 - up your game” DJI [Online] https://www.dji.com/mavic-air-2 Accessed: Dec. 22, 2021

Python, “Welcome to Python.org,” Python [Online], https://www.python.org/ Accessed: Dec. 16, 2021

Jupyter Notebook, “Project jupyter” Juypter Notebook [Online] https://jupyter.org/ Accessed: Dec. 16, 2021]

EA Lab, “NIOSH Sound Level Meter” Apple Store [Online] Available: https://apps.apple.com/us/app/niosh-sound-level-meter/id1096545820 Accessed: Dec. 23, 2021

CDC, “NIOSH Sound Level Meter App,” Center for Disease Control [Online] https://www.cdc.gov/niosh/topics/noise/app.html Accessed: Dec. 23, 2021

Just Warfare: Is a Nuclear Attack an Appropriate Response to a Cyber Attack?

Alexia Fitz and Richard L. Wilson
Towson University, Towson, MD

afitz2@students.towson.edu
wilson@towson.edu

Abstract: It is well known that nuclear weapons pose a grave threat to humanity because of their destructive power and to the lives of innocent civilians potentially affected by them. What might be less known are the grave effects that cyberwarfare can potentially have on a state and its nuclear security. Most nuclear weapons systems were designed decades ago, when manipulations of computer networks, or cyber-attacks, were practically a non-existent threat. In the present international political situation, where threats about the use of strategic nuclear weapons, have been discussed, cyber threats are everywhere, and it may be expected that they will have consequences for the stability of nuclear weapons systems as well. Considering the many unknowns of the continually evolving issues related to cyber threats, it is hard to measure how serious the risks are, but the idea cannot be excluded that, over the long term, they may have "game-changing" effects on the perceived value of nuclear weapons. Potential consequences of this phenomenon include cyber operations targeting nuclear weapons development, nuclear weapons systems, and cyber operations replacing nuclear weapons. It is crucial that nation states, such as for example the United States, work hard to prevent cyber-attacks in a world where we are becoming more reliant on computer systems that if disrupted could destroy the economy, politics, and even military operations. The question, of how far states are willing to go to protect their cyber realm, and the extent to which their nuclear policies might allow for the possibility of a nuclear response to a cyberattack, present significant issues. This analysis will employ distinctions from just war theory, to attempt to address these issues. Just warfare is important to this analysis because developing a sense of right and wrong in a case of unpreventable conflicts could avoid further escalation and even more devastating results. In other words, what occurs when we apply the ethical distinctions of Just War Theory, to Cyber Attacks related to nuclear weapons? Can the distinctions of Just War Theory be employed to create a taboo, so that, the protection of a states cyber security creates an obstacle so that cyber warfare does not lead to a result such as nuclear attack? This analysis also takes into consideration anticipatory research, while developing an argument based on ethical considerations without a nuclear attack having to occur first. Anticipatory research such as this is important as the foundation for developing preventative measures because it can be used to argue for the creation of policies (both domestic and international) that will not allow for a nuclear response to a cyber-attack, therefore eliminating the threat to the international community. This analysis will employ a basic conceptual analysis that will proceed by defining critical terminology and will attempt to address the ethical and anticipated ethical issues related to answering the question, Is a Nuclear Attack ever an appropriate response to a cyberattack?

Keywords: Just Warfare, Cyber Warfare, Nuclear Warfare, Cyber Attack, Nuclear Attack, Appropriate Response.

1. Introduction

The threat of nuclear warfare is not a new concept to the international community. The development of nuclear weapons began taking place in the 1930s when scientists discovered that controlled nuclear fission could be the source of a massive explosion. Since this discovery nuclear weapons have only been detonated twice in times of conflict, both by the United States military in 1945. Currently nuclear weapons have also been developed by ten states, the United States, Russia, China, France, the United Kingdom, Israel, Pakistan, India, North Korea, and South Africa who eventually denuclearized in the 1990s. This means that out of the nearly 194 recognized states in the world only nine states have nuclear weapons. What makes nuclear weapons so unique is their expense and their destructive nature that allows for only a small percentage of states to have access to this technology. These weapons create high levels of international fear due to the possible use of nuclear weapons in the context of warfare. Rapid nuclear weapon expansion took place during the Cold War, well before the creation of modern-day computer systems and cyber threats. However, although nuclear weapons have not been used since World War II the threat of nuclear warfare continues to loom due to inevitable link of nuclear warfare with cyberwarfare.

Cyberwarfare on the other hand is a more recent development that creates fear in the international community as the technology once created for military advantage is now easily accessible to everyone in the political, public and private sectors. Cyber-attacks have continued to increase as cyber technology has continued to develop and pose threats at the individual level all the way up to the government and military. This has led to an increase in security threats within the domain of cyberwarfare which leads to the question, could a nuclear attack ever be an appropriate response to a cyber-attack? This paper will proceed by first exploring the devastating effects

seen with both nuclear and cyber-attacks and will then examine these warfare techniques in the context of just war theory and particularly jus ad bellum criteria. This analysis is focused on applying anticipatory research based on historical nuclear and cyber warfare policies and the current international climate, to suggest preventative measures seeking to protect international security.

2. Background Effects of Nuclear War

The devastating impact and potential effects of nuclear weapons are well known and are major contributing factors which have led to the creation of many international treaties preventing further proliferation of nuclear weapons while promoting the elimination of existing nuclear arsenals. The destructiveness of nuclear weapons has only continued to increase as nuclear weapons technology has advanced and as nation states have competed to gain military advantage over their adversaries. The first ever atomic bomb was tested by the United States while creating an explosion yielding 20 kilograms of force (ICAN n.d.). Less than one month after the first nuclear test, the United States dropped the first ever nuclear bomb on Hiroshima, Japan, on 6 August 1945. The nuclear bomb known as "Little Boy" produced an explosion with 13 kilotons of force killing directly and indirectly (radiation, burns, cancer) over 200,000 people (Atomic Heritage Foundation 2017, "The Manhattan Project"). Then three days later on 9 August 1945 the United States dropped a second nuclear bomb on Nagasaki, Japan. This bomb was given the name of "fat boy" and produced 21 kilotons of force explosion, killing an estimated 74,000 people both directly and indirectly (Atomic Heritage Foundation 2017, "The Manhattan Project"). Since the first use of nuclear weapons, nation states have increased the amount of force that nuclear explosions yield, which could cause even greater casualties and destruction. The newest nuclear weapons developed by states are thermonuclear weapons, also known as hydrogen bombs, produce explosions a 1000 times more powerful than atomic bombs (Chan 2019, "What is the Difference").

Nuclear weapons create powerful explosions capable of destroying structures and killing millions of people, but they also create a several mile wide radii of radioactive debris. The radioactive fallout from a nuclear explosion will have long lasting effects on the surrounding environment even making some areas affected by the blast uninhabitable. If nuclear weapons were to be deployed all of these factors would have tremendous effects on a state's economy, infrastructure, and military operations. Considerable damage can come from a singular nuclear weapon, however nuclear warfare could potentially lead to the use of multiple nuclear weapons from various states, which could potentially result in even more extreme consequences.

3. Effects of Cyber Warfare

Cyberwarfare has become a growing threat as technological advancements extend the threat of cyberattacks in both the private and public sectors. As technology becomes more accessible so do targets for possible cyber-attacks. These cyber-attacks have the capabilities to reach from individuals all the way up to the government, which could have major effects on a state's stability (Atrews 2020). Cyberattacks can come in many different forms such as information breaches, surveillance, and even gaining control of various devices, including power grids, machines, and nuclear weapon systems and controls. For example, many nuclear weapons systems, including those of the United States, are outdated with outmoded cyber security protective measures and as a result, are at risk for potential hacking. If a cyberattack were to occur on a nation states infrastructure it could lead to massive destruction on infrastructure and effect civilian populations, comparable to nuclear attacks including, crumbling an economy, effecting access to electricity and basic needs, and causing political unrest. These effects could eventually lead to a growing number of causalities over time. Similar to nuclear attacks, cyberwarfare comes with the potential for little to no warning before it is too late to prevent the devastation. Some scholars, such as Benjamin Hatch, would go as far as to say the cyber warfare could be classified as a weapon of mass destruction (WMD), as more states have begun to use cyberattacks as an offensive measure in conflicts. Cyber warfare can create destruction, can disrupt the innerworkings of a state, and can create degradation for those affected, all of which fall within the classifications for WMDs (Hatch 2018).

In 2014 alone, the United States government faced up to 60,000 cybersecurity breaches, potentially effecting the lives of millions of people (Atrews 2020). Cyberwarfare poses threats to U. S. and international security and continues to increase risks as technology continues to advance. Many government officials state that cyberwarfare is more of a threat to US national security than any weapons or physical attacks because of how easily accessible technology has become and how reliant people are on that technology (Dever 2013). Unlike nuclear weapons, cyberwarfare involves military strategy and tactics available to a much wider variety of members of the international community and many states have adapted to the threat by establishing their own

departments of cybersecurity within their government systems. In addition, Cyber-attacks are more threatening because they do not necessarily come from enemy nation states, they can also come from civilian attackers with their own individual motivations, while in comparison nuclear attacks involve large scale military operations. According to the Center for Strategic and International Studies, in October 2022 there were eight significant cyber incidents directed against government agencies. For example, Russia in October 2022 accused the United States of cyberattacks against Russia in the U. S. effort to support Ukraine (CSIS 2022). This is a very clear example of how nation state alliances can also be at play in the use of cyberwarfare similar to how members of alliances can support one another related to nuclear warfare strategy. In Ukraine's current conflict with Russia, Ukraine is focusing on traditional warfare strategies but good relations with the US allows the US to get involved in the war, without even setting foot in the warzone. This example also demonstrates the potential consequences of cyberwarfare, which can create increased tensions between nation states with nuclear capabilities. These developments can in addition potentially spark more aggressive military actions, including cyber-attacks that can lead to the threat of the use of strategic nuclear weapons (and in the current situation, the threat of nuclear weapons and retaliation by Russia) (NTI 2022).

4. Just War Theory

The previous section of this paper explored both the similarities and the uniqueness of the destructive capabilities when nuclear warfare and cyber warfare are compared and when they are combined. Next, we introduce Just War Theory, which aims at establishing legitimate criteria for engaging in war according ethical criteria. The criteria in Just War Theory have changed throughout the course of history. Since ancient times scholars have studied the morality of appropriate and inappropriate actions in warfare, discussing the use of force as justifiable if an enemy has engaged in inappropriate behaviour in war? just war theory falls into 4 categories (Internet Encyclopaedia of Philosophy):

> **Jus ad Bellum:** "conditions under which States may resort to war or the use of armed force in general" (ICRC 2015).
>
> **Jus in Bello:** "regulates the conduct of parties engaged in an armed conflict" (ICRC 2015).
>
> **Jus post bellum:** deals with the conditions of states following war including compensation, punishment, and possible rehabilitation (Internet Encyclopaedia of Philosophy n.d.).
>
> **Jus ad vim:** similar to jus ad bellum but it specific to use of force against state and non-state actors, with lesser outcomes compared to traditional [kinetic] warfare (Brunstetter and Braun 2013).

In other words, war is appropriate if used for means of retaliation against acts of aggression by other nation states (Schmitt 2003). The Treaty of Westphalia in 1648 deemed war justifiable if it was in the best interest of the state, essentially arguing that because of sovereignty there was no limit to conflicts. The Hague Conventions also explored the morality of war, justifying war if conflicts were too large to be resolved through mediation. However, the effects of World War I led to stricter limitations of war and the creation of an international judicial system by the League of Nations, which was later replaced by the United Nations and the International Court of Justice, trying to provide states with methods to resolve conflicts other than by war (Schmitt 2003).

Some scholars explore just war theory from a legal perspective where the international system has made the threat of force and use of force illegal (Nussbaum 1943), including modern day developments in the use of nuclear weapons and the use of cyber weapons to wage cyberwarfare. We can conclude that nuclear and cyber warfare are not justified since they are illegal, yet there is a continuation of conflict in today's world. This paper is focusing on just warfare from the perspective of nation states and how they use moral justifications to argue for their continued use of force in war. Jus ad bellum provides the criteria that theorists established to justify when war is permissible by nation states based on ethical considerations. The criteria that make up jus ad bellum analysis include,

just cause: responding to aggressions by an actor, not initiating conflict. In other words, just cause is an act of self-defence. However, "aggression" is nor narrowly defined and does not only have to include physical

violence. Theorists could argue an act of aggression takes form as interfering with a states economy (sanctions), disrespecting a state's culture or practices, or violating a treaty (Internet Encyclopaedia of Philosophy n.d.).

chance of success: a state must weigh the cost and benefits of waging war to determine if reasonable success is possible. There are moral considerations that must go into the calculations, so that monetary and possible casualties are factored in even in the war is won (Internet Encyclopaedia of Philosophy n.d.).

last resort: war should be a final solution only after all other forms of diplomacy and mediation have failed. In all scenarios, war should be a last resort because of the devastating effects it causes to a state's economy and society as a whole (Internet Encyclopaedia of Philosophy n.d.).

proper authority: a government has proper authority because of state sovereignty. The concept of state sovereignty can be a grey area, but theorists have agreed as long as a government has a legitimate ruling over its citizens (therefore making it a proper state) then it has the authority to declare war (Internet Encyclopaedia of Philosophy n.d.).

proper intention: a state engaging in a just war must be doing so with the intentions of seeking peace and justice. A state cannot justify wars for their own national interests or to assert dominance over another state (taking territory, etc.) when there is no longer just cause and it becomes acts of aggression (Internet Encyclopaedia of Philosophy n.d.).

proportionality: is an overlapping concept with jus in bello in the sense that a war must be fought with an end goal. The end goal, however, should not surpass the aggressions that initiated the war in the first place. For example, a state cannot start a just war, but then continue to act aggressively takeover other states land in the process. (Internet Encyclopaedia of Philosophy n.d.).

Each of these criteria provide potential justification for use of force by one nation state against another. Proportionality is an important distinction for addressing the question, "is nuclear warfare an acceptable response to a cyber-attack?" At first the use of nuclear weapons might seem an extreme response to a cyber-attack that involves no direct physical conflict, however the devastating effects of each type of warfare must be considered before answering the question. A difficulty that is immediately present involves responding to a simple issue, whether when a nation state uses cyber-attacks and cyber warfare to attack another nation states nuclear weapon infrastructure, the attacked nation can respond with a nuclear attack. Proportionality refers to the use of violence by one actor which must be comparable to the violence inflicted on them by another actor. Nuclear warfare and cyber warfare can both possibly target millions of military personnel and civilians. Additionally, they both have the ability to inflict extreme damage to a state's economy, political system, and can cause extreme suffering for a nation's population. So, although both cyber warfare and nuclear warfare are illegal in the context of armed conflict, especially when targeting civilians (ICRC 2021), the use of a nuclear weapons as a response to a cyber-attack is a creditable fear depending on the level of destruction that a state sees as a threat.

A brief discussion of the other criteria of jus ad bellum, as they relate to a nuclear response to a cyber-attack, is also useful in determining if this satisfies the criteria of just warfare. Just cause is applicable in the sense that a nuclear attack could be an appropriate response to initial aggressions in the form of a cyber-attack. Chance of success is also plausible for a state choosing to utilize a nuclear response because a state could justify that it saves money and risks high number of casualties by using a nuclear weapon rather than other tradition means of war that can be extremely costly if the conflict is drawn out. The idea of reasonable success is arguably one of the main reasons for the United States' use of nuclear weapons in World War II, which is explained in the next section, "Just Warfare in Practice." Last resort can be argued on a more conditional basis. Depending on the individual situation, a state could argue that other means of diplomacy were utilized, and that a cyber-attack was the final straw in making its decision to use nuclear weapons. Additionally, proper authority is a criteria that currently applies to all international actors that maintain nuclear arsenals because they are all states with sovereignty and legitimate governments. Finally, It could be argued that proper intent is legitimate, on a situational basis, if states retaliate in order to bring peace and not to use nuclear weapons for additional political gains.

A further distinction that needs to be taken into consideration given the current circumstances is Jus ad Vim. In the 2006 edition of *Just and Unjust Wars*, Michael Walzer introduces an important distinction between, "measures short of war," such as imposing no-fly zones, pinpoint air/missile strikes, and CIA operations, and on the other, "actual warfare," typified by a ground invasion or a large-scale bombing campaign. Even if the former are, technically speaking, acts of war according to international law, he argues that "it is common sense to

recognize that they are very different from war." While they all involve "the use of force," Walzer distinguishes between the level of force used: the former, being more limited in scope, lack the "unpredictable and often catastrophic consequences" of a "full-scale attack." (Brunstetter and Braun, 2013)

In terms of state to state conflicts Jus ad Vim distinctions are related to measures that fall short of kinetic acts in traditional conflicts. Cyberattacks that do not cause immediate physical causalities as opposed to nuclear attacks that immediate physical casualties, fall within the criteria that do not legitimate a physical response. The issue that is focused on in this analysis, Is a Nuclear Attack an appropriate Response to a Cyber Attack? needs to address *jus ad vim* acts that are military actions. The moral issues related to *jus ad vim* acts and cyber-attacks and whether they can rise to the level of justifying a nuclear response can be directly related to the proportionality principle in jus ad bellum criteria. it is important to recognize the degree to which non-violent actions such as cyberattacks, align with and adhere to the proportionality criteria in jus ad bellum, while moving beyond *jus ad vim* principles as introduced by Walzer. The important issue that needs be addressed is, could a cyberattack move beyond jus ad vim distinctions to a degree that would trigger a nuclear response in the context that would satisfy the criteria necessary to engaging in just warfare?

5. Just Warfare in Practice

To better understand the argument that nation states could justify a nuclear response to a cyber attack we must look at past uses of just warfare from both a nuclear and a cyber perspective. Anticipatory research must involve examining past and present examples of situations in order to use this material to make projections about potential future outcomes. In the case of nuclear warfare there is only two examples of the use of nuclear weapons in war, the bombings of Hiroshima and Nagasaki. In these cases what prompted the United States' use of nuclear weapons? The decision to drop the atomic bombs on Japan was made by President Truman and his committee of advisers including, Secretary of War, Henry Stimson (Burr 2020). The decision came due to the unwillingness of Japan to surrender, which would lead to the continuation of the war and cause further devastation and human causalities. World War II was one of the most costly and deadly wars and ending it would be beneficial for the United States and the international community. Additionally, Japan had directly attacked the United States territory in it's surprize attack upon Pearl Harbour, killing both civilians and soldiers, which was one of the driving factors for the US to get involved in WWII in the first place. Stimson advocated for not letting the war drag on when the US had the nuclear technology to end it (Burr 2020). President Truman's committee also considered an invasion of Japan or a demonstration of nuclear weapons near Japan, but all other options were considered to be costly and not forceful enough to motivate the Japanese to surrender. The US military believed they had no other choice and decided to bomb Hiroshima because of its importance to the Japanese military (NPR 2015). On August 6, 1945, the US dropped a nuclear weapon on Hiroshima which was followed by another one on Nagasaki on August 9, five days later Japan surrendered. Because the US was the first state to use nuclear weapons in war, they did not break any international laws. Based upon the criteria of just warfare from an ethical perspective the United States' decision to use nuclear weapons came according to the criteria of last resort, proper intention and just cause in order to attempt to end a very costly war. At a fundamental level it can be argued that proportionality was met because the Japanese had targeted US civilians and military bases first, however, it can also be argued that proportionality was violated due to the total destruction and radiation caused by nuclear weapons that was unlike any other weapon used previously in wars.

The destruction caused by the first and only use of nuclear weapons led to the establishment of many international treaties to prevent future nuclear warfare. Treaties like the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), which was open for signature in 1968, was created to prevent an increase of nuclear weapons. NPT only recognized five nuclear weapon states including, China, the United States, Russia, France, and the United Kingdom (United Nations Office for Disarmament Affairs, n.d.). Even with international treaties like NPT in place there is now nine operating nuclear weapons states, Russia, the United States, the United Kingdom, France, China, Israel, North Korea, India, and Pakistan; each with varying policies regarding the use of nuclear weapons. A very common nuclear policy is a "no first use" policy, which ensures that a state will not use its nuclear weapons unless in an act of retaliation against an enemy state that strikes them first (Center for Arms Control n.d.). Currently, India and China are the only countries that have agreed to a no first use policy regarding the use of their nuclear weapons (Global Zero n.d.). India claims to only maintain its nuclear arsenal as a nuclear deterrent against Pakistan, therefore it also agrees to only use its weapons if Pakistan were to initiate nuclear warfare with a nuclear attack. China also has claimed to have a no first use policy, but its creditability has been challenged in recently due to the rapid increase of nuclear weapons in its stockpile and the lack of restraint shown by the Chinese government (Zhao 2021). This leaves a number of key international actors with nuclear

weapons that do not have no first use policies, including the United States, Russia, and North Korea. In theory these states seem to appear to be willing to use their nuclear weapons, if there is a creditable threat to their security due to a cyberattack upon their nuclear weapons system, which goes beyond being a target of a nuclear 1st strike attack. In this case, the destruction could be caused by cyber-attacks could be within the realm of possibilities that could inspire these states to use nuclear warfare in response to a destructive cyber-attack targeting their nuclear arsenal.

6. Current International Climate

Additionally, it is important to examine the current political climate in the international community to understand the use of cyber warfare and the potential for nuclear attacks as well. One of the most notable conflicts as of 2022, is the war between Russia and Ukraine. Russia's invasion of Ukraine that took place on February 24, 2022 and can be seen as a prime example of an unjustifiable starting of conflict by a majority of the international community. Russia's interest in the territory of Ukraine that it claims to be its own, is being used to justify the war and ignore the fact that Ukraine is a recognized sovereign state, which was illegal to invade. In response to Russia's invasion, many states including the United States and its European allies condemned Russia's actions. Ukraine is not a member of NATO; therefore, it is not protected by NATOs nuclear capabilities. However, Ukraine has received aid including conventional weapons and tanks in order to counter Russian attacks. The United States is among many of the countries involved in enforcing sanctions against Russia as a means of punishment for their unlawful war with its neighbouring state. The US and NATOs involvement has led President Vladimir Putin to threaten the use of a first strike and the use of strategic nuclear weapons.

Another rouge state actor contributing to today's highly charged political climate is North Korea under the leadership of Kim Jong Un. North Korea is the latest state to obtain nuclear weapons and has gone against international taboos with the creation of its program and continued nuclear, and ballistic missile testing. The nuclear demonstrations practiced by North Korea has become a form of nuclear terrorism which flaunts their access to deadly weapons in order to deter any potential threats to their sovereignty and security. North Korea continues to be motivated to maintain its nuclear arsenal because of the power all nuclear states hold. For example, a state with nuclear weapons has never been invaded (Ong 2016). Although, this statement is true, the use of cyberwarfare creates a new means of invasion without troops stepping foot onto foreign soil. Any country no matter its military strength, can be affected by cyberwarfare and how they choose to respond to this threat can vary depending on regime type. Any rouge actor, like North Korea, might be quick to respond to a cyberattack with nuclear warfare in order to not appear weak to the international community.

China is another current international actor that should be monitored for both its cyber and nuclear attack capabilities. On one hand, China is extremely well-versed in terms of cyber activities as it has conducted many cyber-attacks for its own national interests. For example, these attacks have targeted the United States government and many other industries including, healthcare, education, energy, and trade, and as a result continues to be a creditable threat to international security (CISA 2022). In addition, the US Defence Department reports that China is rapidly increasing its nuclear weapons arsenal. The report suggests that by 2030 China will have roughly 1,000 nuclear warheads, which is a nearly 300 percent increase of its current stockpile (Bugos 2021). Although China has not openly threatened the use of its nuclear weapons, officials are concerned about China's possible plans of expansion, which seem to include invading Taiwan (Bugos 2021). An invasion of Taiwan would likely result in a similar reaction by the international community to Russia's invasion of Ukraine, This would mean we could see sanctions or cyber-attacks directed at China. Then, depending on how threatened China is by the international response, they could decide to use strategic nuclear weapons.

7. Conclusion

Today's international political climate shows that there is a rise in international tensions from various threats, both from state and non-state actors. The institutions that have been put in place to try to further mitigate deadly conflicts and prevent states from being able to justify wars are arguably becoming less effective. From a realist perspective, just warfare is no longer being considered in conflicts because a state's own interest takes precedence. Technological advances mixed with complicated international relations continue to cause tensions between states and between state and no state actors which increases the possibility of the use of cyberwarfare. Although cyberwarfare is not a conflict in the traditional sense of a physical conflict, it can be perpetuated by both state and non-state actors. To date although there have been a number of incidents that show the potential for inflicting kinetic damage through cyber means (i. e. Stuxnet) no cyber actions have caused physical death. Cyber warfare has the potential when the proper targets are identified, to cause just as much damage to a society as combat warfare. Additionally, states have been more reluctant to implement nuclear policies, such as no first use, and hesitant to adopts treaties that prevent the use of nuclear weapons, like the Treaty on the Prohibition of nuclear weapons (TPNW). Access to nuclear weapons, whether through a state's own arsenal or an ally's, has become a symbol of ultimate security. However, due to the emergence of cyber threats there can no longer be a guarantee of a state's security because even nuclear systems could be a potential target of an adversary's cyberattacks. All of these factors combined could lead to a variety of dangerous outcomes that could potentially include the use of nuclear warfare as a response to a cyber-attack.

It is important to look at this anticipatory research now to better understand the various factors that states might consider amidst a cyber-attack and how to prevent the use of nuclear weapons. Nuclear threats will continue to remain as long as nuclear weapons exist. Therefore, policies must be examined to implement international taboos that make states feel secure and eventually promote nuclear disarmament. For example, the creation of a universal "no-first use" policy would ensure that states will never use their nuclear weapons as a first strike offensive thus reducing the possibility of a nuclear response to a cyber-attack. Perhaps an analogue to this nuclear no first use policy for nuclear weapons needs to be extended and applied to cyberattacks upon adversary's nuclear systems. The no first use policy for nuclear weapons needs to be applied to cyberattacks in order to develop a no first use policy related to cyber- attacks upon an adversary's nuclear systems.

References

Atomic Heritage Foundation. 2017. "The Manhattan Project." *Atomic Heritage Foundation.* https://www.atomicheritage.org/history/manhattan-project (November 15, 2022).

Atrews, RA. 2020. "Cyberwarfare: Threats, Security, Attacks, and Impact." *Journal of Information Warfare* 19. no. 4. 17-28. JSTOR. (November 15, 2022).

Brunstetter, Daniel, and Megan Braun. 2013. *Ethics and International Affairs* 27. no.1. 87-106. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/38F9679507EBFD40E883CFADCC271F27/S0892679412000792a.pdf/from_jus_ad_bellum_to_jus_ad_vim_recalibrating_our_understanding_of_the_moral_use_of_force.pdf (December 7, 2022).

Burr, William. 2020. "The Atomic Bomb and the End of World War II." *National Security Archive.* https://nsarchive.gwu.edu/briefing-book/nuclear-vault/2020-08-04/atomic-bomb-end-world-war-ii (November 15, 2022).

Center for Arms Control and Non-Proliferation. n.d. "No First Use." *Center for Arms Control and Non-Proliferation.* https://armscontrolcenter.org/issues/no-first-use/ (November 12, 2022).

Center for Strategic and International Studies. 2022. "Significant Cyber Incidents." *Center for Strategic and International Studies*. https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents (November 15, 2022).

Chan, Melissa. 2019. "What is the Difference Between a Hydrogen Bomb and an Atomic Bomb?" *Truman Library.* https://www.trumanlibrary.gov/public/2019-10/Development%20of%20the%20Hydrogen%20Bomb-%20Document%20Set.pdf?VersionId=TIums5XoxSxXDVHA.MGW8aqOO6IZItZ2 (November 15, 2022).

Cybersecurity and Infrastructure Security Agency. 2022. "China Cyber Threat Overview and Advisories." *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/uscert/china (December 7, 2022).

Dever, John, and James Dever. 2013. "Cyberwarfare: Attribution, Preemption, and National Self Defense." *Journal of Law and Cyber Warfare* 2. no. 1. 25-63. JSTOR. (November 15, 2022).

Global Zero. n.d. "No First Use FAQs." *Global Zero.* https://www.globalzero.org/no-first-use-faqs/ (November 15, 2022).

Hatch, Benjamin B. 2018. "Defining a Class of Cyber Weapons as WMD: An Examination of the Merits." *Journal of Strategic Security* 11. no.1. 43-61. JSTOR. (December 7, 2022).

International Campaign to Abolish Nuclear Weapons. n.d. "The Road to a World Free of Nuclear Weapons." *International Campaign to Abolish Nuclear Weapons.* https://www.icanw.org/nuclear_weapons_history (November 15, 2022).

International Committee of the Red Cross. 2015. "What are jus ad bellum and jus in bello?" *International Committee of the Red Cross.* https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0 (December 7, 2022).

International Committee of the Red Cross. 2021. "Cyber Warfare: Does International Humanitarian Law Apply?" *International Committee of the Red Cross.* https://www.icrc.org/en/document/cyber-warfare-and-international-humanitarian-law (December 7, 2022).

Internet Encyclopedia of Philosophy. n.d. "Just War Theory." *Internet Encyclopedia of Philosophy.* https://iep.utm.edu/justwar/#H3 (December 7, 2022).

National Public Radio. 2015. "Why Did the U.S. Choose Hiroshima?" *National Public Radio.* https://www.npr.org/2015/08/06/429433621/why-did-the-u-s-choose-hiroshima (November15,2022).

Nuclear Threat Initiative.2020."The Cyber-Nuclear Threat." *Nuclear Threat Initiative.* https://www.nti.org/analysis/articles/cyber/ (December 7, 2022).

Nussbaum, Arthur. 1943. "Just War: A Legal Concept?" *Michigan Law Review* 42. no. 3. 453-479. JSTOR. (December 7, 2022).

Schmitt, Michael N. 2003. "International Law and the Use of Force: The *Jus Ad Bellum." Connections* 2. no. 3. 89-97. JSTOR. (November 15, 2022).

United Nations Office for Disarmament Affairs. n.d. "Treaty on the Non-Proliferation of Nuclear Weapons (NPT)." *United Nations Office for Disarmament Affairs.* https://www.un.org/disarmament/wmd/nuclear/npt/ (December 7, 2022).

Zhao, Tong. 2021. "China and the International Debate on No First Use of Nuclear Weapons." *Asian Security* 18. no.3. 205-213. https://www.tandfonline.com/doi/abs/10.1080/14799855.2021.2015654?src=&journalCode=fasi20 (November 15, 2022).

Walzer, Michael, Just and Unjust Wars: A Moral Argument with Historical Illustrations (New York: Basic Books, 2006 [1977]), pp. xv–xvi Google Scholar.

A Review and Testing of Fault Tolerance Levels of Anti-Poaching Cybersecurity System

Isabelle Heyl, Julia Stone, Takudzwa Vincent Banda, Vian Smit and Dewald Blaauw
Stellenbosch University, Stellenbosch, South Africa

isabelle@nomail.co.za
juliastonee@gmail.com
tadiwanashebanda74@gmail.com
viansmit@gmail.com
dnblaauw@sun.ac.za

Abstract: The development of anti-poaching networks and systems has created a new environment for animals in game reserves all over the world. Advanced technologies such as heat sensors, drones, and trip wires help prevent poachers from entering the property and therefore, creating a safer environment for animals to roam freely. Radio frequency identification (RFID) systems are used to track the location of animals. These networks are, however, susceptible to being hacked if not properly protected with cybersecurity tools, resulting in cyber-criminals gaining access into the network. Many attacks or threats can be executed on the RFID network due to some exposed vulnerabilities of elements within the anti-poaching network. The purpose of this paper is to explore the empirical methods of common attacks, used by cyber-criminals, to attack the anti-poaching network, and whether these methods used are effective in identifying weaknesses within the network. This will be executed by creating an experimental structure of the anti-poaching system with a specialised focus on the RFID elements, using quantitative research methods to produce findings. GNS3, an open-source software application that has specifically been chosen to conduct this research, is used to build the network simulation in order to analyse the weaknesses of the network. Cybersecurity protocols are implemented to protect the network and aim to protect the animals. The attacks performed, such as Flood and Scapy attacks, have shown that the anti-poaching network is vulnerable to penetration from cyber-criminals. A hypothesis test was conducted to determine whether the attacks had a significant effect on the network, by using the average ping time from specific nodes to Google. It was found that the average ping time increased by 2.0020 units, therefore stating that the elements of the network were successfully attacked. The fault tolerance test shows that the availability of the anti-poaching network is roughly 90 percent which concludes that the network is configured to deliver quality performance and handle failures, should there be any intervention. This will allow game reserves to implement and have information on a better and safer RFID system for the animals.

Keywords: Anti-Poaching, Cybersecurity, GNS3, Radio-Frequency Identification, RFID, Cyber-Criminals

Introduction

The illegal activity of wildlife poaching has been active for years, killing innocent animals almost every day to generate money. Poaching statistics for 2022 released by the government of South Africa has shown an increase, and a worrying change in wildlife poaching. Anti-poaching units utilize highly developed technology to protect the wildlife from poachers and to help them fight against these organized crime units. Magnetic sensors are used to capture weapon and gun sounds to locate criminals, as well as a virtual net, which is created around the reserves with networks using the Internet of Things (IoT) to catch and identify poachers rather than animals (Pozniak. H, 2018). Thermal cameras prove to be advantageous as poachers can be seen in low visibility environments, and electric fences are covered with acoustic fibres that alert the staff if it has been tampered with. An active wi-fi connection allows staff to communicate with each other and watch live data while on patrol. Some animals are fitted with tracking devices so their location can always be traced and monitored. These tracking devices, called RFID tags, are connected to an encrypted RFID reader (Pozniak. H, 2018). A RFID network consists of RFID tags and RFID readers which forms part of the anti-poaching cybersecurity system. It is vital to consider the extend of the security of the RFID network in an anti-poaching situation. Anti-poaching units have implemented extensive technology to stand guard against cyber-criminals in the cyber warfare of animal poaching. Therefore, the purpose of this study is to test the fault tolerance levels within an anti-poaching cybersecurity system which is used to protect animals in reserves. This is done by demonstrating a cybersecurity anti-poaching network, performing a critical review of the fault tolerance levels and resulting in an analysis of the system.

Research Question

This paper aims to critically review and test the fault tolerance levels of an antipoaching cybersecurity system.

2.1 Problem Statement

Many high-tech systems are in place to prevent poaching, but unfortunately, they all experience some level of faults. Poachers use this information to gain entry into the anti-poaching system. The system must be able to protect against attacks and safeguard information to maintain efficient functioning. A denial of service (DDOS) attack is mostly performed by cyber-criminals which is executed by multiple packets of information sent to the communication channel thereby increasing the information load and reducing the performance of the reader. Reducing these cyber risks present in this cybersecurity system becomes of utmost importance.

2.2 Aims and Hypothesis

The aim of this paper is to critically review and perform tests to assess the fault tolerance levels within the anti-poaching cybersecurity system. Animal protection units are fighting a cyber warfare against hackers who want to penetrate anti-poaching networks to locate animals as part of their poaching tactic. A hypothesis test is conducted to determine whether the attacks have had a significant effect on the system by testing if the nodes can connect to each other or the internet. This is orchestrated by testing the ping time from each node to the internet, specifically Google, and whether it has increased. This allows us to determine whether our defence mechanisms have helped mitigate the outcome of attacks and protect the safety of animals in the context of poaching.

Literature review

3.1 Relevance of Cybersecurity components for anti-poaching

The importance of wildlife protection is greatly emphasised by both local and global government and wildlife conservation societies to prevent the poaching of animals (Tan et al, 2016). Countries such as, South Africa has passed the Animals Protection Act 71 of 1962, as well as in India, the Wildlife (Protection) Act 53 of 1972 was passed. There are many other laws in place in different countries. These laws were set in place to protect and conserve the ecological and environmental security of all countries. Diminishing wildlife has the possibility of impacting the tourism sector which would put a dent on the nation's long-term environmental and economic prosperity. Wildlife protection will aid the tourism sector (Anderson and Jooste, 2014) which contributed around 3.7 percent of South Africa's GDP in 2021 (South Africa Statista, 2022). Currently, much of the anti-poaching equipment is technologized, and is therefore, considered as major assets of South Africa. RFID systems are one of the main technologies used in anti-poaching systems to conserve wildlife. RFID networks are helpful in protecting the wildlife as they have a constant update on the animal's location. Other components such as, drones, motion sensors and cameras contribute immensely to their different aspects. Hence, maintaining a secure cybersecurity system around the anti-poaching components is fundamental in every game reserve.

3.2 Threats/Attacks to the anti-poaching system

This table describes some of the attacks performed by cyber-criminals when wanting to poach animals.

Table 1: The attacks and threats performed on an anti-poaching system

Type of Attack	Description
Jamming	Preventing communication from the reader to the tag by redirecting the connection to a malicious reader.
Tag cloning	Cybercriminals will snoop the data of the tag and write it to another tag. Thereby, cloning the tag. (Kumar et al, 2021)
Replay attacks	The criminal will duplicate packets and send them to the reader multiple times. The reader then accepts the packets thinking it is the authentic tag.
Passive attacks	Also known as eavesdropping. Hackers gaining an entry into the communication channel by intercepting the signal.

Methodology

5.1 Design proposed model

The systems and technologies used in the RFID reader will be investigated, along with the extent to their cybersecurity. A quantitative approach to research will be used to understand this real-life phenomenon whilst exploring the cyber security aspects of anti-poaching. GNS3 has been chosen as the simulation software because of its many features and vast capabilities. An adapted model is created, using the GNS3 platform, to show the flexibility of the relationship between the RFID reader and the rest of the system, by replicating an anti-poaching cybersecurity system. Kali Linux and various tools will be used through GNS3 to penetrate the simulated RFID system network to expose the weaknesses and threats of the reader as well as the network. The simulated network is an adapted design of a large-scale conventional anti-poaching system. This model, Figure 1, was used to validate the attacks on the network.

The RFID readers are modelled as nodes, while the tags are modelled as disconnected from the system to show how the animals roam around the reserve. The office control room represents a space where the network security management team is based, and where the security technologies are controlled. The room comprises of a network router, switches, RFID servers, management PCs, and a Honeypot. Redundant network design is employed to add additional communication pathways and provide standby routes for when, and if, a connection fails. This is modelled as the standbyRouter, standbyServer, standbyPC, standbyPC2, standbySwitch and loadbalancer. The active system and the standby system are protected with a Cisco Adaptive Security Appliance (ASA) firewall.

IP addresses of all the network nodes are Dynamic Host Configuration Protocol (DHCP) allowing them to use network services such as Domain Name System (DNS), Network Time Protocol (NTP) and Transmission Control Protocol (TCP). The field is a protected area where the wild animals are located and surrounded by wireless connected technological safeguards such as cameras, drones, motion sensors, and smoke detectors. Readers, tags, and backend servers are the three main players in the RFID system.

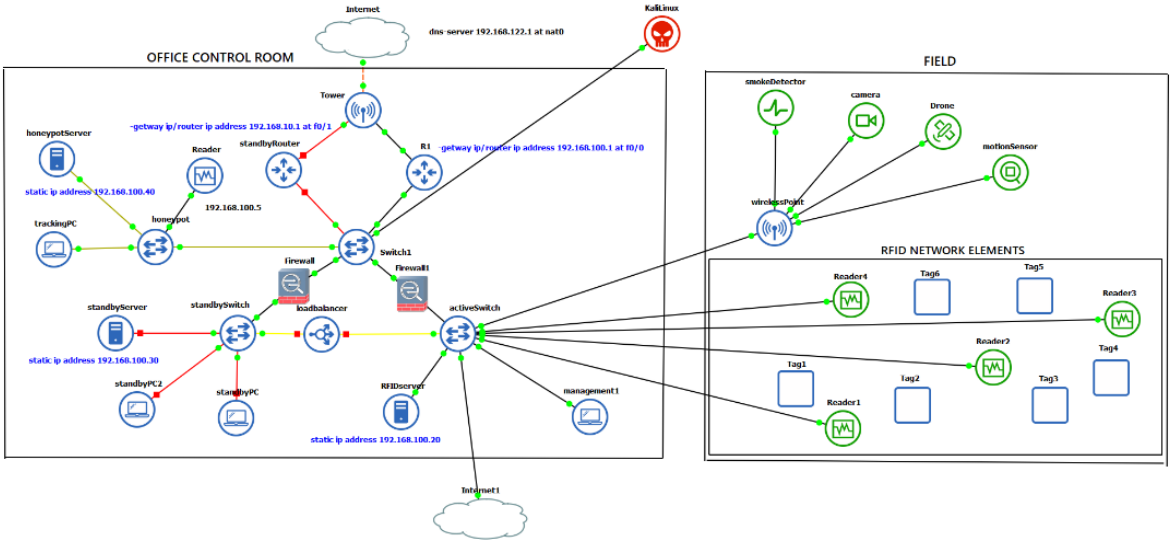


Figure 1: Adapted design of the Anti-poaching system

5.2 Overview of Router configuration

a) Network router (192.168.100.1)

b) Configure IP DHCP pool which assigns and manages IP addresses from network 192.168.100.0/24 within the router to DHCP clients.

c) DNS server (192.168.120.1)

All devices not protected by Cisco ASA firewall are IP DHCP 192.168.100.0 /24 network pool.

All devices protected by Cisco ASA firewall are IP DHCP 192.168.122.0 /24 network pool.

5.3 Implementation

In this phase, Kali Linux will be used to perform network attacks such as, Secure Socket Layer (SSL) STRIP, Address Resolution Protocol (ARP) poisoning, Flood attack, Scapy attack and Macof attack.

5.3..1 SSLSTRIP

SSLSTRIP manipulates the internet traffic of the RFID system, by performing a man-in-the-middle attack. The Tracking PC with IP address 192.168.100.3, was attacked, as shown in Figure 2.



Figure 2: Conducting the SSLSTRIP attack

A Honeypot is set in place to lure network attackers, detect, deflect, and study hacking attempts. The console log did not come up with any data on the login details due to the Honeypot preventing the connection from the attacker to the management pc, as shown in Figure 3. This concludes that the RFID system is secure from cyber-criminals wanting to get access of passwords or sensitive information about the network and bypass the Honeypot.

`

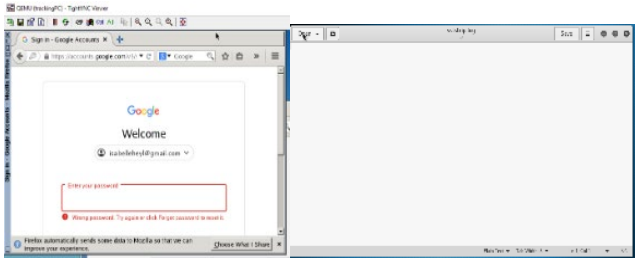


Figure 3: Unsuccessful penetration of the attack

4.3.2 ARP poisoning

The purpose of the Address Resolution Protocol (ARP) attack was to redirect network traffic to the Kali machine (attacker) and therefore, disable the traffic flow for Reader (192.168.100.5). The attack, as shown in Figure 4, was configured to have the default gateway and IP address credentials of the reader.

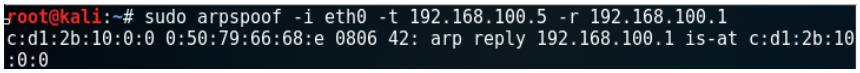


Figure 4: Launching the ARP attack

The Reader was unable to ping Google, as shown in Figure 5, due to Kali (attacker) redirecting the traffic from Reader to the Kali machine. Thereby stating that this attack was successful. The Reader is only protected by the Honeypot and not a firewall which shows that the Honeypot does not protect the network through packet manipulation, thereby Kali has used the Honeypot as an entry point into the network.

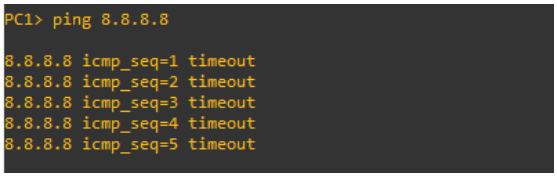


Figure 5: Ping statistics of google from Reader

4.3.3 Flood attack

A very high volume of traffic is sent to the system in a Flood attack which is also known as a Denial of Service (DoS) attack. Figure 6 shows how 15000 packets (-c 15000) at a size of 120 bytes (-d 120) are being sent to port 80 (-p 80). The flood flag (--flood) is being used in conjunction with a transmission control protocol (TCP) sized at 64 (-w 64). The flood attack is executed on the Honeypot (192.168.100.2).

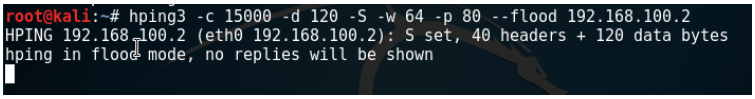


Figure 6: Flood attack on Honeypot

The Management1 node (192.168.100.6) was pinged by the Honeypot (192.168.100.2) as shown in Figure 7, and the destination was unreachable due to previous attacks performed on the Honeypot. This concludes that the flood attack was successful.

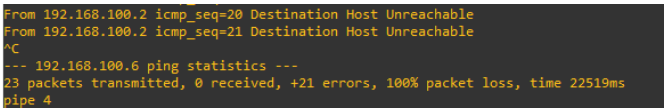


Figure 7: Ping statistics of Honeypot and Management1

When the Honeypot tried to connect to Node (192.168.100.4) it was successful but when Node (192.168.100.4) tried to ping the Honeypot, it was unsuccessful. Figure 8 shows that the Honeypot is still connected to the network yet unable to get a response to the other nodes due to the attack. The Honeypot is protecting the network by rejecting all communication from other nodes.

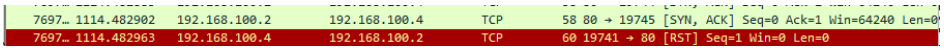


Figure 8: Ping statics on Wireshark

4.3.4 Scapy attack

Scapy is a packet manipulation tool. The goal is to trick Reader3 (192.168.122.212) in the network into sending a legitimate packet to Management1 (192.168.122.181). To do that the attacker will impersonate the reader. Figure 9 shows the execution of the attack.

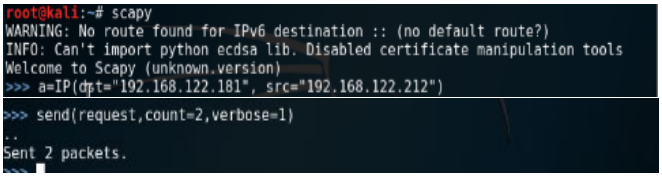


Figure 9: Launching Scapy attack

Figure 10 shows the analysis in Wireshark, this states that the source and destination was computed from Reader3 to Management1, which was completed by Kali, the attacker. This attack was successful, and Kali has broken through Firewall1 due to vulnerabilities present in the firewall.

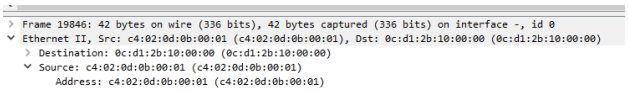


Figure 10: MAC Addresses of Management1 and Reader3

4.3.5 Macof attack

This attack floods a switch with fake mac addresses. The switch doesn't know where exactly to send out the information, so it starts acting like a hub. Anyone listening into the network can then track the information. Figure 11 shows the attack performed on the network.

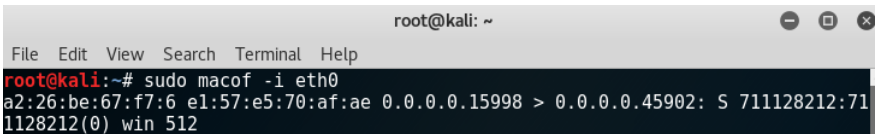


Figure 11: Macof attack executed

Wireshark, in Figure 12, has shown that fake MAC addresses were used as the source of sending many packets of information to the switch. This attack was successful in completing an execution by sending multiple packets and manipulating the switch into thinking they are authentic. This shows that packets are able to move freely around the network.

No.	Time	Source	Destination	Protocol	Length	Info
19782	650.727442	145.225.13.91	110.246.106.66	IPv4	60	
19783	650.727648	222.148.188.78	56.29.209.74	IPv4	60	
19784	650.727876	84.35.96.18	36.82.112.38	IPv4	60	
19785	650.728066	249.69.255.97	133.2.109.89	IPv4	60	
19786	650.733255	31.93.94.66	150.47.237.76	IPv4	60	
19787	650.733553	27.149.112.77	14.96.203.25	IPv4	60	
19788	650.733900	61.89.242.73	102.169.22.121	IPv4	60	
19789	650.734390	97.227.170.52	1.21.22.1	IPv4	60	
19790	650.734699	229.27.15.7	180.55.17.10	IPv4	60	
19791	650.735034	232.67.5.78	121.84.73.108	IPv4	60	

Figure 12: The result shows the fake MAC Addresses

5.4 Attack Summary

These attacks have shown that the RFID system is susceptible to penetration from cyber-criminals and poachers wanting to get information on the animal. Although, if the criminal wanted to get sensitive information of the network system, the system was strong enough to hold, due to the Honeypot preventing the SSLSTRIP attack.

Findings

6.1 Hypothesis testing

R was used to perform an analysis on the datasets. The first step was to determine whether the data are normally distributed. The Shapiro-Wilk normality test was used to test for normality. It was found that the Honeypot data are normally distributed (p-value = 0.614 > α = 0.05), the Tracking PC data are normally distributed (p-value = 0.1115 > α = 0.05), meaning that the majority of the recorded times show similarities. The data of Reader3 are not normally distributed (p-value = 0.002271 < α = 0.05). As a result, we used the paired samples t-test for the Honeypot and Tracking PC datasets, and the Mann-Whitney U test for the Reader3 dataset. The average of five pings to Google per node was taken, before and after an attack. Since the Reader had an output of "timeout" when pinging Google, 0 was used to indicate that nothing happened, as shown in Figure 13. The average ping time for the Honeypot increased by 0.9930 units. The average ping time for the Tracking PC increased by 283.3812 units. The average ping time for the Reader3 increased by 2.0020 units. There was a slight increase in the ping time, as shown in Figure 13.

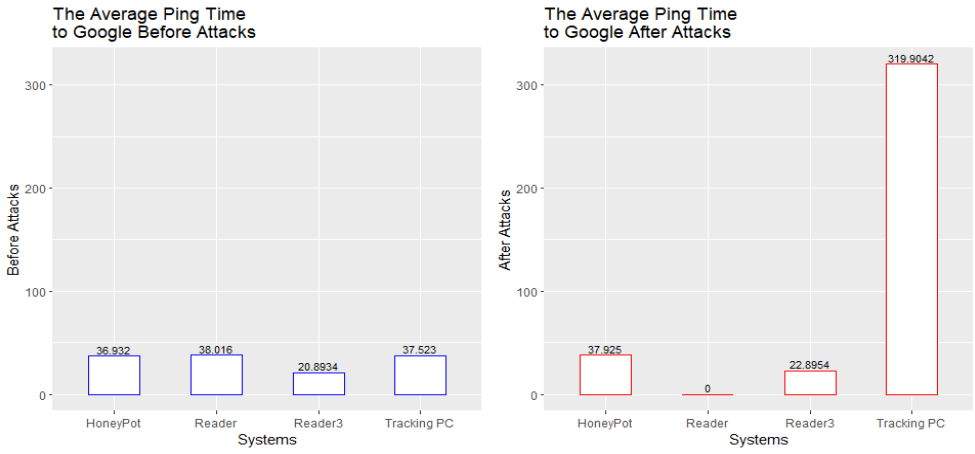


Figure 13: Graphs of the ping-time taken before and after the attacks

5.1.1 Honeypot hypothesis test

Let m = the mean difference in the ping times before and after an attack

$$H_0: m = 0$$

$$H_a: m \neq 0$$

significance level α=0.05

Since the p-value = 0.836 > $\alpha = 0.05$, we do not reject the null hypothesis, we can conclude that the average ping time after the attacks is not significantly different to the average ping time before the attacks.

```
data:  PingTimes by TypeofAttack
t = 0.22087, df = 4, p-value = 0.836
```

Figure 14: p-value for Honeypot

5.1.2 PC hypothesis test

Let m = the mean difference in the ping times before and after an attack

$$H_0: m = 0$$

$$H_a: m \neq 0$$

significance level α=0.05

Since the p-value = 0.0621 > $\alpha = 0.05$, we do not reject the null hypothesis, we can conclude that the average ping time after the attacks is not significantly different to the average ping time before the attacks.

```
data:  PingTimes2 by TypeofAttack2
t = 2.5682, df = 4, p-value = 0.0621
```

Figure 15: p-value for Tracking PC

5.1.3 Reader3 hypothesis test

$$H_0: The\ distribution\ of\ the\ ping\ times\ are\ approximately\ equal$$

$$H_a: The\ distribution\ of\ the\ ping\ times\ are\ not\ equal$$

significance level $\alpha = 0.05$

Since the p-value = 0.4034 > $\alpha = 0.05$, we do not reject the null hypothesis. We can conclude that there is not a significant difference between the distribution of the ping times before attacks and after attacks.

```
W = 17, p-value = 0.4034
alternative hypothesis: true location shift is not equal to 0
```

Figure 16: p-value for Reader3

Even though Figure 13 shows that there is an increase in the ping time after an attack, our hypotheses conclude that there is no significant evidence to support that.

6.2 Fault Tolerance Testing

MTBF (Mean time before fail) = number of hours the system is in use/number of failures encountered

$$MTBF = \frac{28}{2} = 14$$

MTTR (Mean Time Taken to Repair) = number of minutes spent repairing the system/number of repairs

$$MTTR = \frac{3}{2} = 1.5$$

Availability = MTBF/(MTTR + MTBF)

$$Availability = \frac{14}{(14 + 1.5)} = 0.9032$$

We can conclude that the availability of our network is 90.32%

Conclusion

The implementation of cybersecurity within game reserves is desperately needed to protect wildlife and reduce the illegal trade of rhino horns on the black market. Although, this paper has emphasised that attacks on the network can easily be performed, cyber-criminals are proved to not always be successful. This is achieved due to firewalls and Honeypots preventing network penetration, for example, in the SSLSTRIP attack. It is important to note that performing such attacks help in identifying the fault tolerance levels of the RFID system as it pointed out the weaknesses present in the system, and therefore analyses the protection of the network. The fault

tolerance test produced an availability of 90.32 percent which concludes that the anti-poaching network is able to operate continuously at a rate of 90.32 percent without failure. GNS3 is an effective tool in providing the network's security system to be adapted while performing and attempting attacks. The research has concluded that the adapted anti-poaching cybersecurity RFID system is secure and safe from cyber-criminals if proper tools are used to prevent these attacks, such as firewalls. R played an important role in analysing the network while visualising the attack performance through graphs.

Recommendations for Further Research

This paper proposes that additional research must be done to further protect the animals from cyber-criminals. In order to perform a complete hypothesis on whether the ping time increased or not, a larger dataset should be used so that there is a lower estimation variance thus a better predictive model. The results gathered yield sufficient evidence to support the notion that fault tolerance is critical for technological systems. Hardware solutions and software solutions such as the use of Cyber Honeypots, Load Balancing, and the use of firewalls are hereafter recommended for extra protection of the cybersecurity anti-poaching system. A RFID tag could be developed with a larger frequency range to strengthen the connection to the reader, therefore preventing interception from a cyber-criminal. We recommend that further research and more advanced protection tools should be implemented to prevent attackers from gaining access into the RFID network, to further prevent poaching from all aspects.

References

Anderson, B. and Jooste, J. (2014) *Wildlife poaching: Africa's surging trafficking threat*. National Défense Univ Fort Mcnair dc Africa Center for Strategic Studies.

Edwards, J. (2014) *World Wildlife Fund Uses RFID to Foil Poachers-A real-time tracking and monitoring solution helps protect endangered rhinos and other animals in Namibia*. Radio frequency identification (RFID) Journal newsletter.

Ibrahim, A.A.A., Nisar, K., Hzou, Y.K. and Welch, I. (2019) Review and analyzing RFID technology tags and applications. In *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-4). IEEE.

Kumar, A., Jain, A.K. and Dua, M. (2021) *A comprehensive taxonomy of security and privacy issues in RFID*. Complex & Intelligent Systems, 7(3), pp.1327-1347.

Nieto Jiménez, A. and López-Muñoz, F.J. (2019) *GNS3 for Security Practitioners*. Diseño y Configuración de Sistemas Seguros en Red/Design and Configuration of Secure Network Systems.

O'Donoghue, P., and Rutz, C. (2016) Real-time anti-poaching tags could help prevent imminent species extinctions. *Journal Of Applied Ecology*, *53*(1), 5-10. https://doi.org/10.1111/1365-2664.12452

Pozniak, H. (2018) The tech battle we must not lose. *Engineering & Technology*, 13(5), pp.30-35.

Save The Rhino. (2022) *South Africa 2022 poaching stats | Save The Rhino*. [Online] Available at: <https://www.savetherhino.org/member/news/south-africa-2022-poaching-stats/>

Statista. (2022) South Africa: contribution of tourism to GDP | Statista. [Online] Available at: <https://www.statista.com/statistics/1290545/contribution-of-travel-and-tourism-to-gdp-in-southafrica/#:~:text=In%202020%2C%20travel%20and%20tourism,(GDP)%20of%20South%20Africa.>

Tan, T.F., Teoh, S.S., Fow, J.E. and Yen, K.S. (2016) Embedded human detection system based on thermal and infrared sensors for anti-poaching application. In *2016 IEEE Conference on Systems, Process and Control (ICSPC)* (pp. 37-42). IEEE.

DACA: Automated Attack Scenarios and Dataset Generation

Frank Korving and Risto Vaarandi
Centre for Digital Forensics and Cyber Security, Tallinn University of Technology, Tallinn, Estonia

korving.f@gmail.com
risto.vaarandi@taltech.ee

Abstract: Computer networks and systems are under an ever-increasing risk of being attacked and abused. High-quality datasets can assist with in-depth analysis of attack scenarios, improve detection rules, and help educate analysts. However, existing solutions for creating such datasets suffer from a number of drawbacks. First, several solutions are not open source with publicly released implementations or are not vendor neutral. Second, some existing solutions neglect the complexity and variance of specific attack techniques when creating datasets or neglect certain attack types. Third, existing solutions are not fully automating the entire data collection pipeline. This paper presents and discusses the *Dataset Creation and Acquisition Engine* (DACA), a configurable dataset generation testbed, built around commonly used *Infrastructure-as-Code* (IaC) and *DevOps* tooling which can be used to create varied, reproducible datasets in a highly automated fashion. DACA acts as a versatile wrapper around existing virtualization technologies and can be used by blue as well as red teamers alike to run attack scenarios and generate datasets. These in turn can be used for tuning detection rules, for educational purposes or pushed into data processing pipelines for further analysis. To show DACA's effectiveness, DACA is used to create two extensive datasets examining covert DNS Tunnelling activity on which a detailed analysis is performed.

Keywords: security dataset, testbed, DevOps, detection engineering

1. Introduction

DevOps tooling and practices have been widely adopted by both development and IT-operational teams to automate the building, testing and deployment of software and the infrastructure it runs on. One of the effects of this adoption is an increase in trust of the developed solutions since existing *Quality Assurance* (QA) processes can be transformed into more rigid ones by automating and enforcing otherwise labour-intensive steps like regression testing.

Red teams are using these same tools in increasingly creative ways to systematically enforce QA practices like uniquely obfuscating exploit code to avoid being signatured or to test exploit-viability and defence evasion techniques under varying conditions before deployment.

Blue teams can utilize these tools as well by integrating static and dynamic analysers into development pipelines or Vulnerability Management programs. They can also be used to build easily reproducible lab environments which facilitate adversary emulation scenarios and analysis of produced attack artefacts. Security datasets exist that can help with *Intrusion Detection System* (IDS) / *Intrusion Prevention System* (IPS) rule tuning and validation as well as provide valuable practice data for Detection engineers. This has inspired the author to explore the combination of these two topics: automated creation of security datasets using DevOps tooling.

1.1 Motivation

Security Datasets exist in many data types (e.g. Network vs. Flow vs. Log) and data formats (e.g. PCAP vs. EVTX vs. flat file), some are open-sourced, others only available on request, some are artificial, others are captures of a real-world infrastructure setup, some target *Machine Learning* (ML) use-cases while yet others target IDS/IPS use-cases. Tools exist which can manipulate, augment or help create such datasets and can be used to help validate detection capabilities (Brauckhoff et al, 2008; Cordero et al, 2015; Vasilomanolakis et al, 2016).

The creation of such datasets is often a very time-consuming task, however not commonly automated. When tooling *is* created to auto-generate such data it is either not published or shared-on-request, making validation and adjustments very hard.

Unfortunately, published datasets have some other commonly encountered deficiencies like: poor data quality, unclear methodology, are unmaintained or unmaintainable and are often unreproducible (Kenyon et al, 2020).

To illustrate why having high-quality datasets can be important, one can look at a project like Sigma, which aims to provide a vendor-neutral detection format for *Security Information and Event Management* (SIEM) solutions. Of the over 2300 published detection rules that come with the project, only 94 were marked as *stable* as of this writing and were tested with available attack data.

This makes it worth while to investigate whether the creation of a more generic, open-source, automated framework to run adversarial / attack scenarios and extract valuable data is feasible, adjustable and can contribute meaningful datasets to the security community.

1.2 Novelty and contributions

Setting up testbeds using IaC for the use of Security research or dataset creation in particular is not a novel idea on its own. However existing solutions:

- Are not always vendor neutral. Either in the format of the datasets they create, or the technology that is leveraged to create the datasets.
- Are not always open-source or are underdeveloped. Developed tools might be mentioned in a paper, but not publicly released.
- Neglect complexity and/or variance of specific attack techniques when creating datasets.
- Don't fully automate the entire setup, attack, data collection pipeline.

Therefore this research aims to create an open-source, adjustable tool called DACA and/or set of configurations that can run end-to-end automated attack scenarios and extract security datasets from the systems under analysis.

The main considerations that went into the design of DACA are:

- *Modularity* – Many different commonly used provisioning and orchestration tools exist to setup IT infrastructure. The tool should allow for an abstraction layer so that new platforms can be supported with minimal effort or code duplication.
- *Flexibility* – The tool should allow for a wide variety of scenarios and attacks to be carried out where ideally a researcher's imagination is the limiting factor.
- *Functionality* – The tool should be able to create reproducible, useful datasets in a highly automated fashion inspired by *Continuous Integration / Continuous Deployment* (CI/CD) parallelization techniques.

The main contributions of this paper are:

- A publicly available toolkit that creates security datasets for attack scenarios.
- Publicly available security datasets created with the toolkit.
- *Intrusion Detection System* (IDS) rules that have been developed with the help of the datasets.

1.3 Related works

This section covers various testbeds, tools that either generate or manipulate datasets as well as relevant datasets themselves. A more comprehensive overview of these tools and datasets can be found in comprehensive dataset studies of the thesis this work was originally based on (Korving, 2022).

1.3.1 Testbeds

Various testbeds have been created that can facilitate research into information systems or security based research in particular:

- Cloudlab, which provides a cloud-like platform. It can be seen as an extension of Emulab (Hibler et al, 2008), another testbed focused on providing a shared environment for researchers (Duplyakin et al, 2019).
- The DETER Project which is a US national Cyber Security experimentation testbed providing infrastructure and tools to security researchers and educators (Benzel et al, 2011).

The following set of testbeds are not represented in the literature, but do come close to this research's design goals. They are publicly available, create production-like environments with security tooling and are adjustable. However, there is a lack of attack automation and/or automatic data extraction in these testbeds which is a significant drawback.

- Splunk's Attack-Range spins up small lab environments with proprietary detection and analysis solutions, which allows for attack simulation and immediate data analysis for building IDS rules.
- DetectionLab is a set of scripts that allow the automated construction of an Active Directory lab environment with monitoring tooling installed.

One notable domain specific example is an *Industrial Control System* (ICS) targeted testbed called HAI. While the technology stack is not reusable for this particular research, it comes very close to the author's research design and research goals of automating an otherwise very manual attack / dataset extraction process. (Hyeok-Ki et al, 2019)

1.3.1 Datasets

There are multiple relevant public (static) datasets available as well as projects that aim to become security dataset exchanges. Two of the most related to this work are:

- **Security Datasets** (formerly *Mordor*) which is an open-source initiative to collect multi-platform datasets to improve detection capabilities.
- **Canadian Institute for Cybersecurity** associated with the University of New Brunswick produces and publishes many types of security datasets targeting anything from IoT environments to Android malware and *DNS-over-HTTPS* (DoH) (Mohammadreza et al, 2020).

2. Design and Development

DACA is a *Command-line Interface* (CLI)-based tool written in the Python programming language that comes with its own configuration specification to define attack scenarios and acts as a wrapper around existing DevOps tools. For the code that executes scenarios an Object Oriented Programming approach was taken. This allows the tool to be extended to new platforms by implementing an interface (i.e. Controller interface) into new execution classes (e.g. a VagrantController or DockerController class). This avoids having to touch the existing code base and therefore works towards modularity and extensibility design goals of the tool. An orchestration class called ScenarioRunner delegates actual execution to the relevant Controller code.

2.1 Language and Dependencies

DACA was written and tested in Python (v3.8+), dependency management for used libraries is performed using Pipenv. The initial implementation comes only with Vagrant support as a scenario provisioner and therefore needs to be installed together with one of their supported hypervisors (e.g. Virtualbox).

2.2 Execution Phases

When a scenario file is passed to DACA it will go through multiple distinct phases in its execution. An overview of these execution phases can be seen in Figure 1.

1. **Definition** – Scenarios are defined in YAML-based configuration files and can utilize Jinja functionality like variables and filters. These are then transformed into python dictionaries.
2. **Validation** – The python dictionary is validated against a defined schema using Cerberus.
3. **Compilation** – Depending on which backend technology is targeted (e.g. Vagrant) the dictionary is compiled into the appropriate configuration file format describing things like machine type and networking.
4. **Setup** – A provisioning step for the defined machine. Here required services are installed which can be done through inline commands, script files or Ansible playbooks.
5. **Run** – A script or series of commands that should be executed on machine startup. For example the start of a data capture or the execution of an attack-sequence.
6. **Data Collection** – In this step any produced artefacts are collected from the machines in the testbed (e.g. log files and network capture files). Some collection methods require streaming data (e.g. Elasticsearch and Kafka output) and live more in the Setup / Run phases.

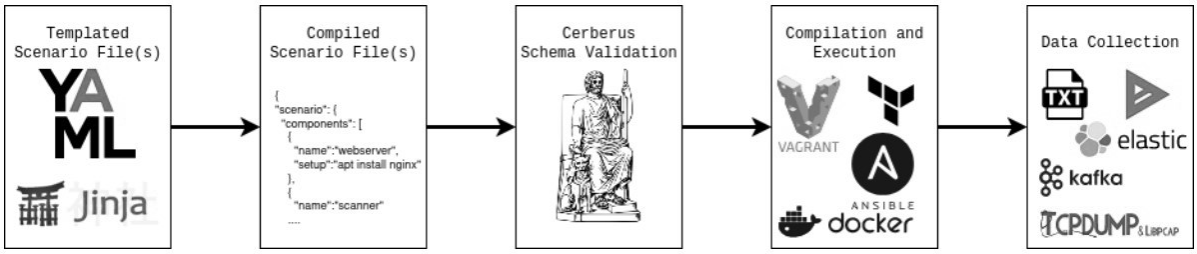


Figure 1: DACA – Step by step execution overview from scenario file to data collection

2.3 Configuration File

DACA supports the definition and execution of attack scenarios through a configuration file. This was implemented by creating a YAML-based configuration language that makes use of a powerful templating engine called Jinja and was used to fulfil flexibility requirements. This solution was chosen in favour of other templating engines because it directly targets python projects and it has proven its value for configuration management tasks by being the primary templating engine supporting some of the most widely used configuration management tools at this time, Ansible and Salt.

To make sure misconfigurations are not allowed to proceed to a Scenario's execution, a schema for the configuration file format was created. This schema definition and validation was performed using a python library called Cerberus. A YAML file is read, transformed into a python dictionary where its keys/values are compared against a schema where attributes like data types, valid values and presence are checked. See the original thesis this paper was based on for the actual DACA configuration file schema (Korving et al, 2022). Cerberus allows to check for slightly more complex, logical errors within the configuration file as well, for example by expressing dependencies between fields or writing custom validation functions. These are not currently used, but could become especially useful when more platforms are supported.

Configuration files have a special section with defined variables and a set of possible values. This section is stripped away in the scenario compilation step after schema validation and substituted into the other sections of the configuration file. This allows a single configuration file to produce multiple execution scenarios. See also Figure 2.

When multiple variables are defined a Cartesian Product is calculated between them, producing a set of variable groups. For example a popular WordPress scanner called "*WPscan*" can enumerate various WordPress objects (e.g. "*u*"-flag indicates user objects) and add Web Application Firewall (WAF) evasion flags. A configuration example for this functionality can be seen in Figure 3 and the resulting set of commands in Table 1. This allows for extensive profiling and possibility of finding unexpected edge cases.

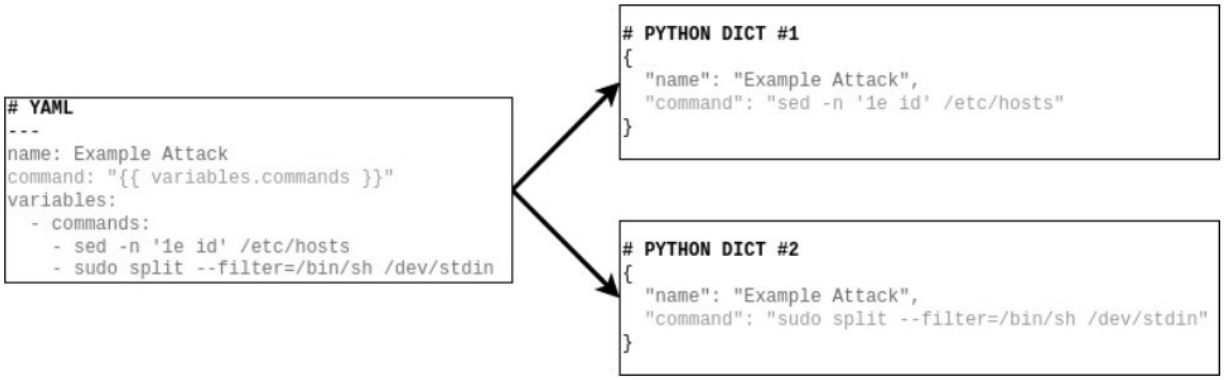


Figure 2: DACA – Single configuration resulting into multiple scenarios using Jinja variables.

```
# YAML
---
name: Example WordPress Attack
command: wpscan --url example.com "{{ variables.flags }}" -e "{{ variables.wp_objects }}"
variables:
  - wp_objects:
    - "u"
    - "cb"
    - "vp"
  - flags:
    - "--stealthy"
    - "--random-user-agent"
```

Figure 3: DACA – WordPress attack variation using Jinja variables.

Table 1: WordPress attack variations

	--stealthy	**--random-user-agent**
u	wpscan –url example.com –stealthy -e u	wpscan –url example.com –random-user-agent -e u
cb	cb wpscan –url example.com –stealthy -e cb	wpscan –url example.com –random-user-agent -e cb
vp	vp wpscan –url example.com –stealthy -e vp	wpscan –url example.com –random-user-agent -e vp

2.4 Compilation

DACA is a wrapper around existing IaC/virtualization technologies which means scenarios need to be translated into a configuration language the target technology understands. This is accomplished using the same Jinja templating engine used in the DACA configuration language itself. The final compiled configuration file is one of the outputs of the tool which make the generated datasets reproducible.

Figure 4 shows a small section of the master Vagrantfile template, its use of variables and basic control structures. Figure 5 shows an example scenario file and the final result. The following 3 variables are used in the template and need to be defined through the DACA scenario file:

- **hostname** – Name of the component.
- **setup['type']** – This variable controls how the value stored in *setup['val']* is interpreted. '*shell*' forces inline command execution while '*script*' and '*ansible*' expect a filename for a shell script or Ansible playbook, which is copied to the target machine and then executed during runtime.
- **setup['val']** – The inline commands, shell script name or Ansible playbook filename which will be executed.

```
1  # Provision VM
2  {%- if setup['type'] == 'ansible' %}
3  {{ hostname }}.vm.provision "ansible_local" do |a|
4      a.install = true,
5      a.install_mode = "default",
6      a.playbook = '{{ setup['val'] }}',
7      a.become_user = "root",
8      a.become = true
9  end
10 {% elif setup['type'] == 'shell' %}
11 {{ hostname }}.vm.provision "shell", inline: '{{ setup['val'] }}', privileged: true
12 {% elif setup['type'] == 'script' %}
13 {{ hostname }}.vm.provision "shell", path: '{{ setup['val'] }}', privileged: true
14 {% endif %}
```

Figure 4: DACA – Templated provisioning section in Vagrantfile.

```
# YAML
---
name: bind9
description: Example BIND9 DNS Server Provisioning
provisioner: vagrant
setup:
  type: ansible
  val:  bind9_playbook
```

→

```
# VAGRANTFILE
bind9.vm.provision "ansible_local" do |a|
    a.install = true,
    a.install_mode = "default",
    a.playbook = 'bind9_playbook',
    a.become_user = "root",
    a.become = true
end
```

Figure 5: DACA – Scenario to Vagrantfile compilation

2.5 Data Collection

DACA allows for multiple types of data to be collected by specifying an "*artefacts_to_collect*"-section under a defined component in a scenario. The currently supported outputs include:

- **Files** – A list of files can be specified which will be collected when the scenario run-through has ended.

- **PCAP** – This expects a tcpdump command which is run as a background task on startup and cleanly stopped at shutdown. Any produced PCAP files need to be specified in the "*files*"-section for them to be collected.
- **Filebeat** – A list of files can be specified to be monitored using Filebeat. Output will be directed to "*/tmp/filebeat.json*" and can be more useful than simple flat files since a lot of metadata will be attached that can be retroactively ingested into Elasticsearch for easy searching. Specifying a list of files under the "*Filebeat*", the "*Elasticsearch*" or the "*Kafka*"-section triggers a dedicated Ansible playbook to be run for installation and configuration of Filebeat. Since a Filebeat instance only supports a single output, these three options are mutually exclusive.
- **Elasticsearch** – Expects an Elasticsearch HTTP endpoint. All files defined in the "*files*"-section (except for PCAP and shell recording files) are monitored and pushed to the defined endpoint.
- **Kafka** – Expects the address of a Kafka listener. Pushes data into a dedicated topic per monitored logfile. This output can be useful when live data streams or additional post processing is wanted (e.g. data normalization).
- **Asciinema** – Asciinema is a terminal session recording utility which allows for attacker-perspective recording and playback. This stanza requires the name of an output file. It will cause the shell commands that are part of the "*run*"-section (and their outputs) to be recorded.

3. Validation

3.1 Implementation Comparison

As mentioned in Section 1.3.1, multiple other testbeds exist. A limited comparison between these available testbeds can be seen in Table 2, highlighting some of the most important qualities in relation to this work.

Aspects like scalability, usability and flexibility are important but are left out since they are subjective or hard to quantify without performing an extensive comparative analysis. While DACA has only limited platform support as of this writing, its focus on automation makes it stand out when compared to the other identified testbeds.

Table 2: Implementation Comparison.

Tool	Publicly Available	Proprietary Data	Supported Platforms	Automated
DACA	Yes	No	Linux, Partial Windows through Vagrant	Fully, Partial
Splunk Attack Range	Yes	Yes	Linux, Windows, AWS, Azure through Vagrant and Terraform	Partial
DETER	No	No	Linux, Windows, Containers through QEMU and Openvz	Partial
Cloudlab	No	No	Linux, Windows through OpenStack	Partial
Detectionlab	Yes	No	Linux, Windows, MacOS, AWS, Azure and more through Terraform	Partial

3.2 DNS Tunnelling Scenario

The following section discusses the run-through of one of the implemented DNS tunnelling scenarios which functions as a test and validation of the developed tool. In total 136 attack variations were performed each resulting in their own sub dataset. 116 were gathered in fully automated mode (file transfer) and 20 of these were performed in DACA's interactive mode (C2). The evaluated DNS servers were: PDNS, Bind9, DNSmasq and CoreDNS. The evaluated tunnelling tools were: Iodine (file transfer only), dns2tcp and dnscat. Suricata was used for additional data analysis.

To design a scenario which emulates a DNS Tunnelling attack it needs to be boiled down to the minimal needed components while still being able to generate the data samples one wants to obtain. The three needed components are: a "compromised" client, a victim DNS resolver and a malicious authoritative DNS server. These components are all deployed within the same network, where requests for a faux domain (*example.attack*) are directly forwarded from the provisioned victim DNS servers to our provisioned attacker authoritative DNS server. Figure 6 shows a step-by-step execution overview of these attacks:

1. The scenario gets rendered into multiple instances, each of which gets compiled into a valid Vagrantfile. Attacks vary by the used tunnelling tool (e.g. iodine vs. dnscat2), data encoding (base32 vs. base64), record type (e.g. TXT vs. NULL) or victim DNS server (e.g. Bind9 vs. CoreDNS).
2. The wrapper-class around the vagrant CLI-utility brings all the defined Virtual Machines (VM) up.

3. The provisioning sections are executed which install the DNS servers and setup zone forwarding / logging configurations. It also prepares the attack server and client.
4. Run-time triggers are defined to start DNS packet captures, start the DNS servers and initiate the DNS tunnel.
5. Once the main attacker's run-time command exits, the VMs are shutdown which triggers the shutdown actions. This includes stopping the packet capture, collecting files and performing artefact clean-up tasks.
6. Data is collected and stored in a dedicated path, along with metadata before moving on to the next execution.

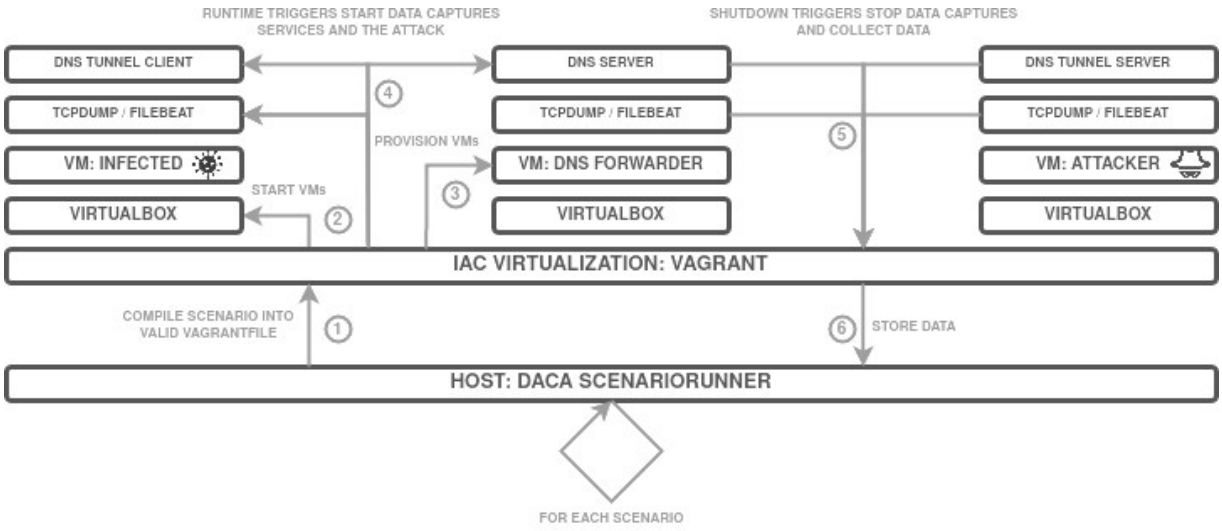


Figure 6: DACA – ScenarioRunner / VagrantController Scenario execution.

3.3 Data Analysis

This section provides a short analysis of data collected during scenario executions, for more details please see the original thesis this paper was based on (Korving et al, 2022). The analysis of collected data allows to assess how tested solutions are responding to attacks and what evidence they produce about the attacks. Furthermore, the evidence that has been identified during data analysis allows human experts to write signatures for detecting these attacks in production environments.

To give an idea of the volume of log data that gets generated when running six exploratory C2 commands ("*whoami*", "*w*", "*pwd*", "*uname -a*", "*ip a*", "*env*"), on average approximately 250 log lines were created for the tunnel establishment, control messages and commands/responses. There is a small but noticeable difference between the two DNS tunnelling tools used: dnscat2 and dns2tcp. It also reveals the fact that DNS servers can produce wildly different amounts of log volumes when query logging has been enabled. PowerDNS needs mentioning since it produced an order of magnitude more data than any of the other tested DNS resolvers although it bears mentioning that the solution's documentation warns users of this fact.

An additional volumetric difference between the DNS Servers was that Dnsmasq had more than double the amount of log lines than the other two remaining DNS servers under evaluation: BIND9 and CoreDNS. While all tested DNS servers log incoming queries, only two make attempts at logging responses, these latter two do not, which explains the difference in volume. The lack of response logging diminishes forensic usefulness since C2 commands and/or any dropped payloads are contained within the responses.

When looking at the produced data itself we can make some other observations, starting with how DNS servers handle writing logs to disk when special or non-printable characters are present in the request/response. These characters mostly occur due to query compression to increase link bandwidth but other mentionable encoding issues occur as well. We would expect similar characters to occur when for example other injection attacks are being performed (Philipp et al, 2021). Octal escape of the raw byte seems to be the standard way to approach this problem (i.e. *000* to *255*) and is adopted by CoreDNS, PowerDNS and Bind9 while Suricata converts special characters to Hex with double backslashes (i.e. "*xee*" or "*xc8*"). This escaping of non-ASCII characters is prudent to avoid injection attacks. Dnsmasq's approach was found to be somewhat inconsistent. No user queries are logged when even a single non-ASCII character is found. However dnsmasq seems to allow logging the

original query in the response message when all characters fall within the wider ISO 8859-1 character set. The response will be printed character by character until a prohibited character is encountered and seems to abandon response logging altogether (even without the 'unprintable' message) when a non supported byte or record type is encountered. See also Figure 7.

```
1 | Mar 31 07:16:42 dnsmasq[1586]: query[TXT] <name unprintable> from 192.168.0.30
2 | Mar 31 07:16:42 dnsmasq[1586]: forwarded <name unprintable> to 192.168.0.20
3 | Mar 31 07:16:42 dnsmasq[1586]: reply rayadíÙBûüý½« REST OMITTED
          »ü.example.attack is r << REMNANT RESPONSE DATA IS NOT PRINTED >>
```

Figure 7: Special characters in log lines.

One basic way to creating signatures for these tools is to find cleartext markers or control messages used by the tool when establishing the tunnel. Figure 8 shows some example ones used by dns2tcp and Figure 9 shows some for iodine.

```
1 | XXXXX.=auth.example.attack
2 | XXXXX.=connect.example.attack
3 | XXXXX.0k.example.attack
4 | XXXXX0.k.example.attack
```

Figure 8: Dns2tcp control messages.

```
1 | XXXXaAbBcCdDeEfFgGhHiIjJkKlLmMnNoOpPqQrRsStTuUvVwWxXyYzZ.example.attack
2 | XXXXaA-Aaahhh-Drink-mal-ein-J\228germeister-.example.attack
3 | XXXXaA-La-fl\251te-na\239ve-fran\231aise-est-retir\233-\224-Cr\232te.example.attack
```

Figure 9: Iodine control messages.

As part of this analysis we've seen a lack of response logging in BIND9 and CoreDNS, excessive debug logging in PowerDNS and inconsistent handling of non-printable characters in Dnsmasq. Some straightforward way to profile these tools was discussed. In addition query logging can have significant overhead to the DNS server and induce query response latency. If one wants to reliably detect the described or similar techniques some other solutions exist that could be considered:

- Dnstap is natively supported by BIND9, CoreDNS and PowerDNS (amongst others) and can be used to avoid IO performance penalties by using a binary output format as well as gain response logging. Data forwarding might be a problem with this solution making centralized analysis harder.
- Tcpdump or DNSCAP can collect DNS traffic passively which has potentially similar data collection issues as Dnstap but would give the full exchange.
- An IDS like Suricata or network analysis tool like Arkime can ingest network data through network TAPs and allows for request/response DNS logging and analysis.

3.4 Evaluation Discussions

3.4.1 Limitations

At the moment the tool only supports the design of local VM based scenarios, which means cloud-native attack scenarios are not yet supported nor can the cloud be utilized to scale out scenarios or parallelize the scenario execution phase. This also means it's not yet possible to design scenarios built onto other virtualization platforms like Docker. Adding support for Docker would not add the ability for new types of scenarios to be developed, in fact it's not suitable when analysing a class of attack scenarios where privileged access is needed to the underlying host (e.g. when installing/reconfiguring an EDR or HIDS solution). However for many situations it would greatly shorten the execution time and would allow more parallelization which in turn allows for a shorter scenario design-development-test lifecycle.

Other limitations of the tool exist that are related to networking and Windows platform support. This is addressed more in the original work (Korving et al, 2022).

3.4.1 Availability information

The code of DACA has been published under a permissive open source license and can be found at the following link: https://github.com/Korving-F/DACA/

Two example DNS tunnelling datasets described in this paper were similarly released and can be found here:

- https://github.com/Korving-F/dns-tunnel-dataset
- https://github.com/Korving-F/doh-tunnel-dataset

4. Future Work

The presented work can be extended, matured and improved upon. The following list describes some ideas on how to do so. Extending to the cloud and other virtualization platforms were already mentioned, so were some limitations that can be addressed.

- Windows Domain – A common requirement for production-like scenarios would likely be to have all participating machines be domain-joined. This requires specialized knowledge on the part of a scenario designer, not relevant to the scenario design at hand and would add a lot of boilerplate code to many scenarios. This setup should be templated for simple deployments and be enabled through a flag in the scenario configuration file (i.e. "*create_domain: true*").
- Built-in Analytics – Live data analysis is dependent on externally setup Kafka and Elasticsearch clusters and their network being reachable. It would be a nice thing to also include these solutions, as well as a data normalization pipeline from within the testbed itself.
- Another challenging research topic is the creation of benign datasets. DACA supports the creation of such datasets, but requires extensive time and energy to be invested into the scenario design.
- Since one of the main value propositions for DACA is automation, scenarios built around automatable C2 (e.g. Mythic) and Adversary Emulation tools (e.g. Atomic Red Team) is something that the author would like to see implemented.

References

Benzel, T. (2011) *The Science of Cyber Security Experimentation: The DETER Project*, Proceedings of the 27th Annual Computer Security Applications Conference, pp 137–148.

Brauckhoff, D. and Wagner, A. and May, M. (2008) *FLAME: A Flow-Level Anomaly Modeling Engine,* USENIX CSET '08

Cordero, C. and Vasilomanolakis, E. and Milanov, N. et al. (2015) *ID2T: A DIY dataset creation toolkit for Intrusion Detection Systems*, 2015 IEEE Conference on Communications and Network Security (CNS)

Duplyakin, D. and Ricci, R. and Maricq, A. et al. (2019) *The Design and Operation of Cloudlab*, USENIX ATC '19.

Hibler, M. and Ricci, R. and Stoller, L. et al. (2008) *Large-Scale Virtualization in the Emulab Network Testbed*, USENIX 2008 Annual Technical Conference

Hyeok-Ki, S. and Woomyo, L. and Jeong-Han, Y. et al. (2019) *Implementation of Programmable CPS Testbed for Anomaly Detection*, 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19),

Kenyon, A. and Deka, L. and Elizondo, D. (2020) *Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets*, Computers & Security, Vol. 99.

Korving, F. (2022) *DACA: Automated attack scenarios and dataset generation*, Master thesis, Tallinn University of Technology, Department of Software Science.

Mohammadreza, M. and Logan, D. and Gurdip, K. and Arash, H. (2020) *Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic*, The 5th IEEE Cyber Science and Technology Congress, Calgary, Canada, August 2020.

Philipp, J. and Haya, S. (2021) *Injection Attacks Reloaded: Tunnelling Malicious Payloads over DNS*, 30th USENIX Security Symposium (USENIX Security 21).

Vasilomanolakis, E. and Cordero, C. and Milanov, N. (2016) *Towards the creation of synthetic, yet realistic, intrusion detection datasets*, 2016 IEEE/IFIP Network Operations and Management Symposium

Towards a Critical Review of Cybersecurity Risks in Anti-Poaching Systems

Christelle Steyn and Dewald Blaauw
Centre for AI Research (CAIR), School of Data Science & Computational Thinking, and Department of Information Science, Stellenbosch University, Stellenbosch, South Africa

xtel911@gmail.com
dnblaauw@sun.ac.za

Abstract: Anti-poaching operations increasingly make use of a wide variety of technology for intelligence and communications. These technologies introduce cybersecurity risk, and they need to be secured to provide greater protection to the information and people involved in anti-poaching operations, ultimately protecting vulnerable animals better. A hypothetical network of anti-poaching technologies was simulated in Graphical Network Simulator 3 (GNS3), consisting of various field devices identified in the literature, and a main control room with relevant hardware devices. A virtual Kali Linux machine was connected to the network and played the role of a digital attacker or intruder. Several cyber-attacks were carried out, to show the risks inherent to such an interoperable and socio-technical network. These attacks included Man in the Middle (MitM) and Denial of Service (DoS) attacks. These attacks were then mitigated via system configurations. Further risks and threat considerations were identified in the literature. Using the STRIDE, DREAD and Attack Tree threat models, the risks to an anti-poaching network were classified and calculated. The most prevalent threats and the attacks performed in the simulation were all calculated to have a high risk level, posing a great threat to an unsecured network. The STRIDE classes of Denial of Service and Elevation of Privilege posed the most risk to the system, both having a calculated average risk score of 9 out of 10. Mitigations to general network threats and those identified in the simulation are mentioned. Additionally, authentication for such a system was investigated, as improper authentication practices were deemed a risk and provides a foothold for further risks in the network. Recommendations made, include the proper configuration of network devices, especially the router and switch, and the use of anti-virus, firewalls, and intrusion detection systems, as well as having an external audit performed annually. Multi-factor authentication, with a password/fingerprint combination, is recommended.

Keywords: cybersecurity, anti-poaching, GNS3, threat model

1. Introduction

1.1 Background Information

Poaching is the exploitation of wildlife that a state considers to be illegal (Duffy, 1999). Poaching is problematic for wildlife conservation (Cooney et al, 2016), and since 2008, poaching has reached a crisis point. Anti-poaching refers to the measures put in place to counter poaching. To combat poaching, conservationists make use of different technologies for intelligence and communication, and this has improved anti-poaching efforts (Thevar and Bhanot, 2021). Utilising technology introduces cybersecurity risks to a system and requires that a system is adequately protected and securely authenticated. Thevar and Bhanot (2021) state that poachers may have access to advanced technology that can infiltrate or attack anti-poaching systems.

Cybersecurity measures need to be implemented on socio-technical anti-poaching systems, which contain interoperable technologies and human actors. "Understanding, managing, controlling and mitigating risk" to relevant critical assets, is Sotnikov's (2022) definition of cybersecurity.

1.2 Problem Statement

Anti-poaching systems contain sensitive data, and a reliance on increasingly accessible technology, make them vulnerable to exploitation by progressively sophisticated attackers (Schoenfield, 2015). The formulation of risk mitigation strategies should have a high priority, as the poaching crisis is still a current problem, and the security of people and animals need to be guaranteed. While studies have been conducted on the use of various technologies in anti-poaching systems, very few address the specific cybersecurity concerns of such systems.

1.3 Limitations

It was anticipated that information regarding anti-poaching operations may be difficult to find, due to their confidential nature. The consultation of various sources can aid in forming a clear composite picture of a hypothetical, interoperable anti-poaching system.

1.4 Research Questions

The following research questions were asked:

RQ1: Can cybersecurity risks be adequately identified within a hypothetical anti-poaching system and mitigated using simulation tools?

RQ2: Can the risks identified in an anti-poaching system be calculated with threat models?

Together with a subsidiary question:

SQ1: What cybersecurity countermeasures are currently in place to protect the anti-poaching industry and what additional countermeasures can be added or modified to improve the safety of animals?

2. Literature Review

Literature on individual anti-poaching technologies is easy to source, but literature is scarce on current anti-poaching systems, and scarcer regarding the cybersecurity of these systems. By isolating the researched technologies and their relevant cybersecurity concerns, a picture of the risks faced by anti-poaching systems can be filled in. Academic sources, South African government guidelines and private initiatives' data were consulted.

Anti-poaching technologies identified to be pertinent to this study are:

- CCTV (DEA, 2020)
- Camera traps (CCF, 2022)
- Sensors (Yayha et al, 2019)
- Drones (Chapman & White, 2019)
- Mobile Apps (CCF, 2022)
- Tags (Bridge et al, 2019)

Common cyber-attacks on systems include Denial of Service (DoS) and Man in the Middle (MitM) attacks. These attacks can render systems unusable or infiltrate them to access information.

The then South African Department of Environmental Affairs (DEA) issued a guideline for anti-poaching systems, containing a threat management framework and suggested technologies (DEA, 2020). Another document contains a strategy regarding the creation of a national information system where rhino data can be centralised to advise security decisions, and calls for regular risk assessments (DEA, 2010).

The Connected Conservation Foundation (CCF), a technology backed anti-poaching initiative, was founded in 2015 by Dimension Data and Cisco (CCF, 2022). CCF makes use of numerous interoperable technologies to blanket a reserve and has succeeded in reducing poaching incidents (BusinessWire, 2018).

Borges (2021) lists malware (viruses, rogue software, trojans, spyware, worms), DoS attacks, phishing, rootkits, SQL injections and MitM attacks as the most prevalent threats to a system. Naagas and Palaoag (2018) compiled a comprehensive list of threats to an information system that can be applied here. Whitman (2004) also provides a list of threats and additionally names protection mechanisms that were employed – of which 100% of survey respondents used passwords. Authentication and identity management are important considerations for the security of systems.

Risk revolves around the probability of a threat affecting an asset, and the consequences of this. There are various types of risk, but here the focus is on cybersecurity risks. Threat modelling is a methodology where anticipated risks are modelled, to aid in determining countermeasures (Gonzales, 2022). This study will focus on several threat models, to aid in calculating risk.

Gonzalez (2022) states that pre-empting, isolating and countering threats can be achieved with threat modelling, which is vital to the organisation's security posture. There exist many threat models and threat models are not always used in isolation and combining models can provide a more holistic picture of the threat landscape. Attack Trees will be used to showcase overviews of existing threats. For the most prevalent system threats, STRIDE will be used to classify threats and the risk will be calculated using DREAD.

Authentication mechanisms make it more difficult for an attacker to breach a system and its failure poses a risk to a system. Passwords are a fundamental part of authentication but are easily disclosed by cyber-attacks (Almehmadi & Alsolami, 2019). The authors conclude that implementing multi-factor authentication (MFA) is necessary to deal with password vulnerabilities.

3. Method

3.1 Methodology

A hypothetical anti-poaching system was simulated in GNS3, running on a VMWare virtual machine, and can be seen in Figure 1. A virtual Kali Linux machine was connected to the system to simulate the role of an attacker.

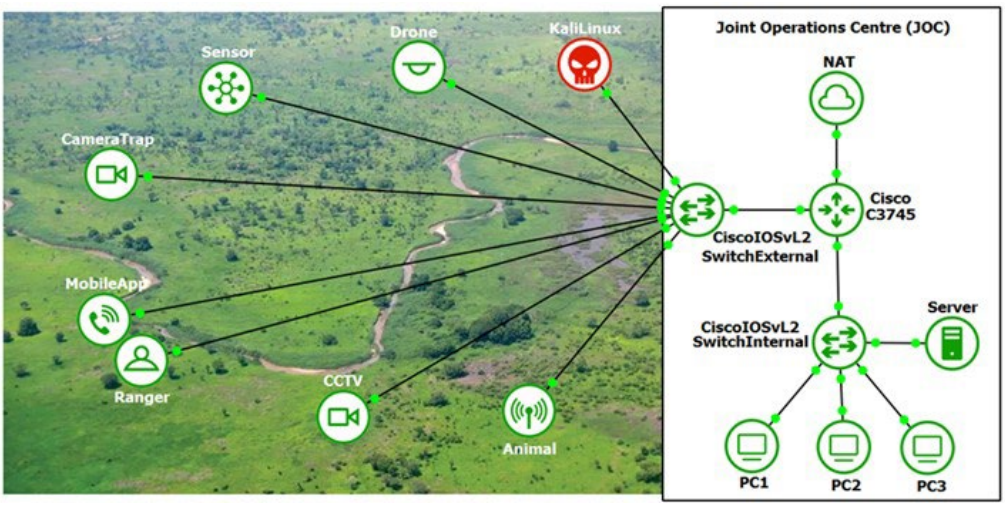


Figure 1: A hypothetical anti-poaching system simulated in GNS3

The Joint Operation Control Centre (JOC) is the headquarters of an antipoaching operation. All components are connected here via a switch and router to the Internet and to each other. PC1-3 represents the monitoring devices employed in the JOC.

The network was configured with a Dynamic Host Configuration Protocol (DHCP), and it was verified that all components received individual IP addresses and could communicate with the network and Internet.

3.2 Cyber Attacks

ARP poisoning, DHCP spoofing and starvation, STP attack, TCP/SYN flooding, CAM overflow and VLAN hopping were the attacks deployed on the network. These attacks were chosen because they are common attacks, cover different attack types (such as DoS and MitM attacks), and are considered among the common threats to a network by Borges (2021), and have been successfully simulated before (Bergs et al, 2022)

3.2.1 ARP Poisoning

IP forwarding was enabled on the attacking machine and Ettercap was used to launch the attack. The ARP of CameraTrap (IP 10.1.2.104) was poisoned. The attack was successful, as the component could not ping Google (8.8.8.8) and PC1 (10.1.1.101).

To mitigate this attack, DHCP snooping, and ARP inspection was applied to VLAN (Virtual Local Area Network) 1 on SwitchExternal (seen in Figure 2). Redeploying the attack, the console provided alerts that the mitigations were active, but the switch was unreachable from CameraTrap. The cause of this is uncertain and could be attributed to a software flaw.

```
Switch(config)#ip dhcp snooping vlan 1
Switch(config)#ip arp inspection vlan 1
```

Figure 2: Mitigations applied against ARP poisoning

3.2.1 DHCP Spoofing

MobileApp's (10.1.2.103) IP was cleared and verified to be 0.0.0.0. The attack was launched by providing Ettercap with a spoofed IP pool, Netmask and DNS server, as seen in Figure 3. MobileApp was instructed to reacquire its IP, however it acquired the IP from the legitimate DHCP server instead of the spoofed one.

DHCP Snooping can be applied to prevent DHCP spoofing, but its success could not be verified in this case.

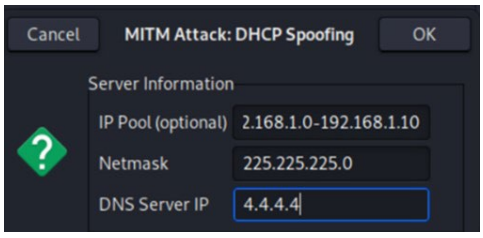


Figure 3: DHCP spoofing attack setup in Ettercap

3.2.1 DHCP Starvation

In the Kali machine, Yersinia was utilised to deploy this attack. This attack should slow the functioning of SwitchExternal, and Sensor (10.1.2.105) was used to verify this. Sensor was only able to intermittently ping Google and PC1, with every other ping experiencing a timeout, meaning the attack was successful.

DHCP snooping and rate limiting were used to counter the attack. The rate limit was set to 100 on SwitchExternal's interface with the Kali machine, as per Cisco (n.d.)'s recommendations (seen in Figure 4). When the attack was redeployed, Sensor was able to ping Google and PC1, making the mitigations effective.

```
Switch#
Switch#configure t
Enter configuration commands, one per line.  End with CNTL/Z.
Switch(config)#ip dhcp snooping vlan 1
Switch(config)#interface g0/1
Switch(config-if)#ip dhcp snooping limit rate 100
```

Figure 4: Setting DHCP snooping and rate limits on switch

3.2.1 STP Attack

The Spanning Tree Protocol (STP) of VLAN 1 on SwitchExternal was viewed, with note being taken of the root address. Yersinia was instructed to claim the root role of the switch, which resulted in the root address being altered (Figure 5a and 5b) and the port now showing it to be the that of the Kali machine.

```
Switch>show spanning-tree vlan 1

VLAN0001
  Spanning tree enabled protocol ieee
  Root ID     Priority    32769
              Address     0cb7.b0c5.0000
              This bridge is the root
```

5a

```
Switch>show spanning-tree vlan 1

VLAN0001
  Spanning tree enabled protocol ieee
  Root ID     Priority    32769
              Address     0cb7.b0c4.0000
              Cost        4
              Port        2 (GigabitEthernet0/1)
              Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec
```

5b

Figure 5: a) SwitchExternal's root address before the attack b) Altered root address and port after the attack

The main mitigations for the STP attack are to apply portfast, bpduguard and guard root. Once these have been implemented and Yersinia relaunched its attack, it could be seen that the mitigations proved successful, as the root role of the switch was not altered.

3.2.1 TCP/SYN Flood

Hping3 was commanded to flood SwitchExternal (Figure 7), with the intention of monitoring network traffic with Wireshark. Here the computer resources proved insufficient to make use of Wireshark while the attack was active and had to be abandoned. Ranger (10.1.2.102) was inspected to view the effects of the attack. Pinging Google resulted in timeouts or high latencies, but pinging PC1 or the switch was possible with normal latencies.

```
┌──(root💀kali)-[~]
└─# hping3 -V -c 100000 -d 10000 -S --flood 10.1.2.1
```

Figure 7: Command to launch TCP/SYN flood via hping3

A firewall was implemented by Bergs et al (2022) to combat this attack but could not be implemented here or be verified to be effective.

3.2.1 CAM Overflow

The MAC (Media Access Control) table of SwitchExternal was inspected and showed that it had ample space left for new connections (Figure 8). Using the macof tool on the Kali machine, it continuously generated and sent MAC addresses to the switch. The expectation was to see that the MAC table now had zero space left, but instead it was found that SwitchExternal went down and ceased to respond to commands. Animal (10.1.2.100) was used to ping Google and PC1. The former experienced high latency and timeouts, and the latter presented with intermittent high latency.

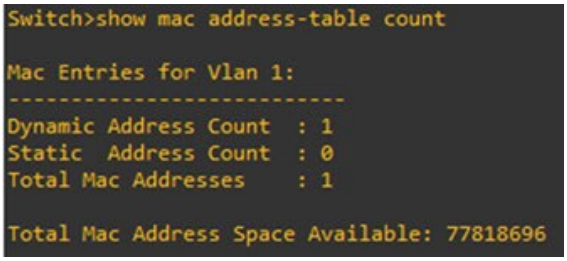


Figure 8: SwitchExternal's MAC address table, showing space available

Port security for the Kali machine's port was enabled on SwitchExternal to mitigate the attack. Heintzkill (2021) recommends only enabling five MAC addresses to be received from the Kali machine, after which the port will shut. SwitchExternal then withstood a relaunched attack and Animal could ping Google and PC1 with normal latency and no interruptions.

3.2.1 VLAN Hopping

After verifying that no trunks were present, Yersinia was used to launch an attack where it "enables trunking" on SwitchExternal. Upon reinspection it could be seen that the Kali Machine had created a trunk link on the switch, which could enable the attacker to access traffic on the link.

To mitigate this attack, trunking was disabled, and DTP (Dynamic Trunking Protocol) prevention was enabled on the switch, seen in Figure 9. After attacking again, the switch was prompted several times for a trunk link, but none were present, making the mitigation successful.

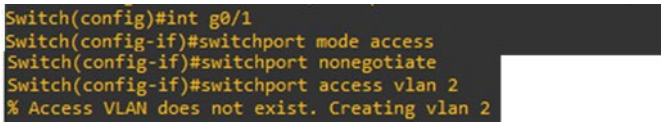


Figure 9: Mitigating VLAN hopping on SwitchExternal

3.3 Threat Modelling

Three threat models were considered for this study. Attack Trees are diagrams that include a root node as the attacking goal, with leaf nodes showing the ways in which an attacker can achieve this goal (Schneier, 1999). The STRIDE threat model can be used to classify threats into six categories – Spoofing, tampering, repudiation, information disclosure, denial of service or elevation of privilege (Abomhara, Gerdes & Køien, 2015). The DREAD threat model assigns numerical scores to identified threats and then translates theses scores to a qualitative risk level (Zhang et al, 2021). DREAD is short for:

- Damage Potential: How much damage can a threat cause?
- Reproducibility: Can the threat be easily duplicated by others?
- Exploitability: How easy is it to exploit the threat?
- Affected Users: Internally or externally, how many users will be impacted by the threat?
- Discoverability: Can one easily detect the threat?

3.4 Authentication

CERN (2018) provides a comprehensive list of good password practices and recommendations that can be considered, as it could be assumed that passwords will be present in an anti-poaching system. To satisfy MFA, additional authentication mechanism needs to be considered. Ali, Dida and Sam (2020) investigated the types of authentications most effective against certain threats, seen in Table 1.

Table 1: Authentication countermeasures against threats (Ali, Dida, and Sam, 2020))

Threat	Authentication Countermeasure
Spoofing	Biometric
Phishing	MFA
Trojan	Biometric
MitM attack	Biometric (fingerprint) and MFA
DoS attack	Biometric (fingerprint)

The use of password managers can also be considered, as CERN (2018) state they are effective in creating and storing strong passwords, as well as verifying previously accessed websites for legitimacy

4. Analysis and Findings

4.1 Cyber-attacks Summary

Table 2 provides a summary of the cyber-attacks performed, categorising them as MitM, DoS or Layer 2 (L2) attacks. It shows that several attacks can affect an anti-poaching network and that risk is evident. However, the table also indicates that these attacks can be mitigated with appropriate network configurations and component implementations.

Table 2: Cyber-attacks summary

Type	Attack	Effect on Network	Mitigation	Result
MitM	ARP Poisoning	Pings were unsuccessful and suffered timeouts	DHCP snooping and ARP inspection applied to switch	Switch not reachable, but alerts appear indicating mitigations
MitM	DHCP Spoofing	No effect, failed attack.	DHCP snooping (not applied here for this attack)	N/A
DoS	DHCP Starvation	Pings suffered higher latency or timeouts	DHCP snooping and rate-limiting applied to switch	Enabled the switch to limit and filter traffic coming from the Kali machine
MitM	STP Attack	The Kali machine assumed the root role of the switch	portfast, bpduguard and rootguard applied to switch	Ensured continued authority of the switch and disabled the Kali machine's ability to interfere
DoS	TCP/SYN Flood	Pings were intermittently successful, but suffered timeouts and high latency	Firewall (Not applied here)	N/A
DoS	CAM Overflow	Pings suffered occasional timeouts and higher latency	Activated port security on the Kali machine's port	Pings were successful with normal latency
L2	VLAN Hopping	Kali machine creates a trunk link and can access network traffic	Disabled DTP and trunking	No trunking was detected

4.2 Threat Model

An attack tree was created, in Figure 10, envisioning the paths a cyber-attacker might take to breach the JOC, based on the attack presented here. The tree shows the goal of the attacker, to poach an animal, and the leaf nodes show how they intend to achieve this. "Break In" was greyed out as, in the scenario, it is deemed to be impossible for the attacker to achieve this and they must therefore go via the virtual route.

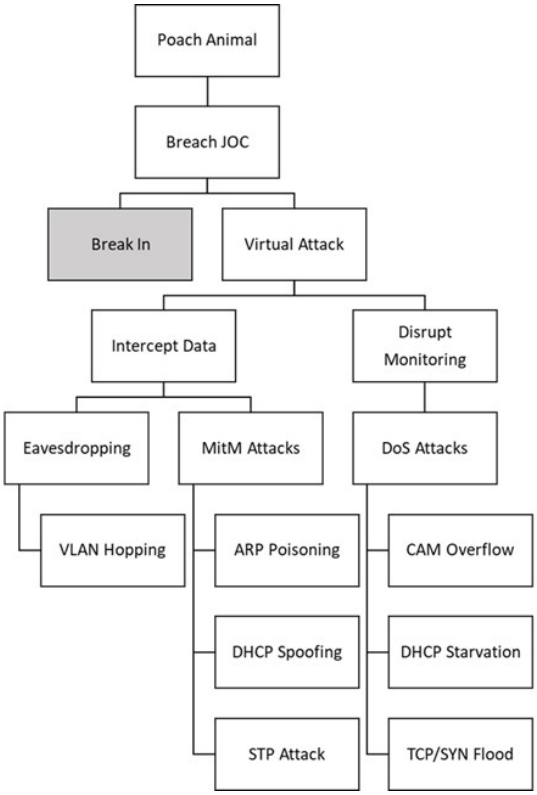


Figure 10: Poach Animal attack tree

Naagas and Palaoag (2018) provide the following equation to calculate DREAD risk scores: $rs_t=(D+R+E+A+D)/5$. A risk level of 7-10 is considered high, while 4-6 is medium and 1-3 is low.

First each DREAD class is individually scored and then all are added and averaged to get the final risk scores. Naagas and Palaoag (2018) investigated an information system and scored its risks, and their scores have been applied to the attacks and threats mentioned in this paper, in Table 3.

Table 3: STRIDE/DREAD threat and risk score table

Threat	STRIDE	D	R	E	A	D	Risk	Level
Malware								*High*
-Viruses	T	10	10	10	10	10	10	High
-Rogue Software	S	10	10	5	5	5	7	High
-Trojans	S	10	10	5	5	5	7	High
-Spyware	I	10	10	5	5	5	7	High
-Worms	T	10	10	5	5	5	7	High
DoS Attacks								*High*
-DHCP Starvation	D	10	10	5	10	10	9	High
-TCP/SYN Flood	D	10	10	5	10	10	9	High
-CAM Overflow	D	10	10	5	10	10	9	High
MitM Attacks								*High*
-ARP Poisoning	S	10	10	5	5	5	7	High
-DHCP Spoofing	S	10	10	5	5	5	7	High
-STP Attack	E	10	10	5	10	5	8	High
Phishing	I	10	10	5	10	5	8	High
Rootkits	E	10	10	10	10	10	10	High
SQL Injections	T	10	10	5	10	5	8	High
VLAN Hopping	I	10	10	5	10	5	8	High

From the above table, viruses and rootkits seem to pose the most individual risk to a system. The individual STRIDE classes can be viewed to find the specific class with the highest risk. Table 4 helps to determine that the class of Denial of Service and Elevation of Privilege are tied at 9/10 for highest risk. With regards to these classes, Naagas and Palaoag (2018) recommend Intrusion Detection Systems (IDS) to combat DoS attacks, and Maayan (2021) recommends a reliable antivirus, regular monitoring, the avoidance of phishing and up to date software as mitigations for rootkits, the main attack vector here for Elevation of Privilege. Due to the nature of the simulation, attacks falling under the Repudiation class could not be performed here.

Table 4: STRIDE risk scores and levels

STRIDE	Total Risk Score	Number of Threats	Average Risk Score	Average Risk Level
S	28	4	7	High
T	25	3	8.33	High
R	-	-	-	-
I	23	3	7.67	High
D	27	3	9	High
E	18	2	9	High

4.3 Multi-Factor Authentication and Authorisation

A password will be employed for secure access to an anti-poaching system. For additional security, a biometric mechanism can be added, with a fingerprint appearing to be the best option. A geographical location can be added to the MFA, in the case of secure data retrieval and storage inside the JOC. The access of peripheral agents, that often change location, to sensitive data can be verified by an OTP (One Time PIN) mechanism. While OTPs themselves can be vulnerable to attack, Peeters et al (2022) state that OTPs can be very effective in practice and that many users prefer it over other authentication mechanisms.

5. Conclusion

A hypothetical anti-poaching system can be simulated and attacked, mostly via MitM and DoS attacks. Mitigations can also be successfully applied.

With the gathered data, threat modelling took place. Attack trees captured visual depictions of threat sources and scenarios, while the STRIDE and DREAD models were used to classify and calculate risk. Individually, viruses and rootkits appear to pose significant risk to a network, but as a STRIDE class, Denial of Service (D) and Elevation of Privilege (E) maintained the highest average risk levels.

If authentication is securely implemented, with password managers and MFA, it lessens the risk for cyber-attacks to gain access to the network, as certain authentication mechanisms have been proven to remain resilient against attacks such as MitM and DoS.

5.1 Research Answers

RQ1: Yes. Risks were adequately identified with literature, simulation, and threat modelling.

A vast quantity of academic resources is available on information system risks, as well as cybersecurity countermeasures. The literature provided informed on risks, and these were incorporated into a simulation and threat model.

Cyber-attacks were successfully performed in the simulation, proving that risk is inherent in an unconfigured anti-poaching system. Only seven cyber-attacks were performed in the simulation, mostly DoS and MitM attacks, as other types of attacks were not always possible, such as password cracking or virus deployment. The simulation software here provided most of the capabilities with which one could mitigate these attacks.

The literature on threat models aided in identifying the risks associated with information systems. The threat models classified and calculated risks.

RQ2: Yes. The STRIDE and DREAD threat models were employed. The DREAD model proved useful in quantitatively scoring and calculating risk, and then translating it into qualitative terms based on the risk score. Classifying a risk with the STRIDE model and scoring it with the DREAD model ensures that risks can be prioritised and that the most problematic threats are highlighted.

SQ1: This question cannot be answered with the available data. While sources exist on anti-poaching technologies, no sources were found on what cybersecurity countermeasures are currently in place on such systems. No person working with these systems was found to be available for comment, even though contact was made with several organisations actively involved in this. Thus, additions and modifications cannot be made in confidence. Recommendations can be made, but will assume an unprotected system, and these recommendations will be general in nature or specific to the risks identified here.

5.2 Recommendations

The key recommendations that stand out as necessities when setting up an anti-poaching system, are:

- Install a comprehensive and reputable antivirus
- Keep software up to date
- Use Intrusion Detection Systems (IDS)
- Implement a firewall with appropriate configurations
- Educate personnel and create security awareness
- Have an expertly and securely configured network, router, and switch configuration.

This is not an exhaustive list and does not guarantee total security. It is recommended that at least one well-rounded cybersecurity expert is employed in anti-poaching operations, that can manage system risks and adapt the system when necessary. It can be investigated whether it may be feasible and beneficial for an independent cybersecurity audit to be performed annually.

While not performed here, regular risk assessments can complement threat modelling and be compiled into a risk framework, to guide anti-poaching cybersecurity strategies.

Regarding authentication, MFA is considered essential, where the combination of passwords and fingerprint biometrics are recommended. For certain use cases, geographical location factors or OTPs can be incorporated. Passwords should be strong and unique, and organisations can customise their password guidelines according to best security practices. Using an enterprise password manager should be considered, as this can generate and store highly secure passwords, avoiding the risk of insecure password behaviour.

5.3 Future Work

To accurately test the robustness of an anti-poaching system, an existing physical system would need to be assessed. New information can be gathered via surveys, interviews, and observations. More sophisticated cybersecurity tools can be introduced to protect the system and the effects of honeypots, DMZs, firewalls, and sniffers could be observed.

It can be investigated whether there is a more suitable threat model to apply to anti-poaching systems.

System authentication was reviewed, and research can be done into creating a more robust authentication system, via access control, MFA, and cryptography, that is practical for anti-poaching systems.

References

Abomhara, M., Gerdes, M. & Køien, G. M., 2015. A STRIDE-Based Threat Model for Telehealth Systems. *Norsk informasjonssikkerhetskonferanse* (NISK2015), pp. 1-15.

Almehmadi, T. & Alsolami, F., 2019. Password Security in Organizations: User Attitudes and Behaviors Regarding Password Strength. *16th International Conference on Information Technology - New Generations* (ITNG), Volume 800, pp. 9-13.

Bergs, C.-J., Bruiners, J., Fakier, F. & Stofile, L., 2021. Cyber Security and Wind Energy: A Fault-Tolerance Analysis of DDoS Attacks. International Conference on Cyber Warfare and Security, pp. 1-13.

Borges, E., 2021. Top 10 Common Network Security Threats Explained. [Online] https://securitytrails.com/blog/top-10-common-network-security-threats-explained [Accessed 30 June 2022].

Connected Conservation, 2022. Technology Solutions. https://connectedconservation.foundation/technology/ [30 January 2022].

Bridge, A. S. et al., 2019. An Arduino-Based RFID Platform for Animal Research. Frontiers in Ecology and Evolution, 7(257), pp. 1-10.

BusinessWire, 2018. Dimension Data and Cisco Take Anti-Poaching Technology into Africa. https://www.businesswire.com/news/home/20180508006383/en/Dimension-Data-and-Cisco-Take-Anti-Poaching-Technology-into-Africa [30 January 2022].

CERN, 2018. Password Recommendations.https://security.web.cern.ch/recommendations/en/passwords.shtml [28 July 2022].

Chapman, L. A. & White, P. C. L., 2019. Anti-poaching strategies employed by private rhino owners in South Africa. Pachyderm, Volume 61, pp. 179-183.

Cisco, n.d. Configuring DHCP Snooping. https://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst6500/ios/12-2SXF/native/configuration/guide/swcg/snoodhcp.pdf [17 May 2022].

Department of Environmental Affairs, 2010. National Strategy For The Safety And Security Of Rhinoceros Populations In South Africa. http://www.stoprhinopoaching.com/wp-content/uploads/2021/08/DEA-National-Rhino-Strategy.pdf [3 February 2022].

Department of Environmental Affairs, 2020. Guidelines to Inform the Establishment of Anti-Poaching Related Systems and Services. https://www.environment.gov.za/sites/default/files/legislations/guidelinesforantipoaching_systemsestablishment.pdf [16 October 2021].

Gonzalez, C., 2022. Top 8 Threat Modeling Methodologies and Techniques. https://www.exabeam.com/information-security/threat-modeling/#:~:text=There%20are%20six%20main%20methodologies,threats%20facing%20your%20IT%20assets. [10 June 2022].

Heintzkill, R., 2021. CAM Table Overflow Attack Explained. https://www.cbtnuggets.com/blog/technology/networking/cam-table-overflow-attack-explained [18 May 2022].

Maayan, G., 2021. How to prevent a rootkit attack. https://blog.malwarebytes.com/how-tos-2/2020/01/how-to-prevent-a-rootkit-attack/#:~:text=To%20fully%20protect%20yourself%20against,then%20reinstall%20the%20entire%20system.&text=Phishing%20is%20a%20type%20of,or%20downloading%20an%20infected%20attachment.[6 July 2022].

Naagas, M. A. & Palaoag, T. D., 2018. A Threat-Driven Approach to Modeling a Campus Network Security. *ICCBN 2018: Proceedings of the 6th International Conference on Communications and Broadband Networking*, pp. 6-12.

Peeters, C., Patton, C., Munyaka, I., Olszewski, D., Shrimpton, T. & Traynor, P., 2022. SMS OTP Security (SOS): Hardening SMS-Based Two Factor Authentication. *ASIA CCS '22: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security,* pp. 2-16.

Schneier, B., 1999. Attack Trees. https://www.schneier.com/academic/archives/1999/12/attack_trees.html [6 June 2022].

Schoenfield, B. S. E., 2015. Securing Systems: Applied Security Architecture and Threat Models. Boca Raton: CRC Press.

Sotnikov, I., 2022. How to Perform IT Risk Assessment. https://blog.netwrix.com/2018/01/16/how-to-perform-it-risk-assessment/ [25 May 2022].

Thevar, S. & Bhanot, N., 2021. How Technology has improved Anti-poaching Efforts. https://sanctuarynaturefoundation.org/article/how-technology-has-improved-anti-poaching-efforts [8 April 2022].

Whitman, M. E., 2004. In defense of the realm: understanding the threats to information security. *International Journal of Information Management*, Volume 24, pp. 43-57.

Yahya, A., Bogaisang, K., Gamoshe, O. G. & Maina, M. D., 2019. Anti-poaching System using Wireless Sensors Network. BIUST Research and Innovation Symposium, pp. 1-3.

Zhang, L. et al., 2021. A risk-level assessment system based on the STRIDE/DREAD model for digital data marketplaces. *International Journal of Information Security*, Volume 21, pp. 509-525.

Work in Progress Papers

Locality-based Electromagnetic Leakage Assessment Using CNN

Ian Heffron and James Dean
Air Force Institute of Technology, Wright-Patterson AFB (Air Force Base), Ohio, USA

ian.heffron@afit.edu
james.dean@afit.edu

Abstract: Deep learning side-channel analysis (SCA) attacks have recently gained in popularity. The ability of deep learning models to retrieve a key byte while minimizing the need for pre-processing steps in both cases of misaligned traces and when the leakage model is multivariate has contributed to the popularity gain. Near-field electromagnetic (EM) probes have enabled leakage capture with high spatial resolution, and the field of deep learning side-channel research looks to find models which reduce the required number of leakage traces to retrieve a key byte successfully. However, despite the many papers researching techniques to reduce the number of traces required for a successful attack, location-based research for EM SCA remains untouched. Due to the nature of EM probes and the architecture of different boards and chips, the location of the collection probe becomes important when attempting to extract the secret key within a reasonable timeframe and with a level of certainty in the result. Our contribution is a framework to determine the best location to assess localized EM leakage against a given chip platform. We use a raster scan to collect localized leakage results at 25 points in a 5x5 grid pattern on a ChipWhisperer Lite XMEGA board running an open-source implementation of the AES-128 encryption algorithm. We demonstrate the use of our framework to locate the best points for an attacker to execute a profiling attack by identifying the grid point with the most detectable leakage for a chip platform. We further execute a smaller localized 25-point raster scan in a grid over the identified location to further refine our estimate of the optimal collection location. We then demonstrate that this location can be used to execute a side-channel profiling attack against the same chip architecture and will result in a lower number of traces required to retrieve the key byte.

Keywords: side-channel analysis, localized leakage, neural networks, electromagnetic leakage, leakage detection

1. Introduction

Side-channel analysis attacks are a class of attacks that seek to exploit the physical phenomena related to cryptographic primitives. During any encryption operation, sensitive variables are processed using a combination of a known public variable, called plaintext, and a secret variable, a key. Due to processor architecture and the manufacturing process, information leakage occurs during encryption operations. This information leakage comes in the form of EM, power, timing, and other physical phenomena. These phenomena can be observed and recorded to perform SCA work and SCA attacks. Encryption algorithms such as the Advanced Encryption Standard (AES) can be defeated by SCA profiling attacks. Profiling attacks involve collecting profiling traces on a prototype device, similar to the intended target device. The profile traces are then used to train a neural network to create a model that will be used in the attack. Attack traces are then gathered on the intended target device and predictions are generated based upon the attack traces and the generated model. A Bayesian process is then used by the attacker to determine the likelihood for each potential key byte value. The candidate byte value with the highest probability is predicted to be the correct key (Masure, Dumas and Prouff, 2019). These profiling attacks are often performed one key byte at a time in a divide-and-conquer strategy. If each byte can be retrieved, the whole key can be reassembled. Much of current SCA research focuses on retrieving the correct key bytes while minimizing number of required attack traces. This paper seeks to further this focus and presents a framework with which to determine, for a given chip, a point with the highest perceived leakage from which to collect leakage and execute an attack. Simply put our hypothesis is that there are points on a device that exhibit leakage that is more susceptible to an electromagnetic SCA attack than other points on the device.

1.1 Our Contribution

We develop and demonstrate the effectiveness of a framework designed to determine the highest perceived leakage point for an attacker to collect EM leakage information for a given chip and board combination. We do this by training a neural network to detect leakage in traces collected at several points above the target. At each point we collect training and testing traces where both the plaintext and key values are randomized. The attack traces are collected using randomized plaintext, and a randomly generated but fixed key. We train the network using a combined dataset of all training traces collected at each of our points. Then we evaluate the accuracy of our model by feeding the fixed number of testing traces from each point into our model and generate a heatmap based upon the returned accuracies. The points on the heatmap with the highest scores indicate areas with the strongest leakage. We then perform attacks using the traces collected at the points with the strongest and weakest leakages to demonstrate that the key byte can be retrieved more reliably and more quickly at the point with the strongest leakage. Using this framework, we can demonstrate that a single trained network given

information from all collection points can correctly ascertain the best point on a given architecture to retrieve key byte information when fed into a neural network. We do not contend that this deep learning architecture is the optimal architecture for this problem, however we are able to demonstrate the effectiveness of the technique to find points with higher discernable leakage.

2. Background

This section will briefly discuss the terminology and background material required to understand our paper. We will discuss previous research, deep learning, and SCA attacks.

2.1 Previous Research

The work of Schlösser et al, (2013) successfully demonstrates identification of the exact SRAM location accessed during the activation of the substitution box (SBox) lookup table during AES encryption operations. Andrikos et al., (2019) uses neural networks and an information-based approach to retrieve information about the AES encryption key after finding the SRAM location of the lookup table containing the key by using the framework given in (Schlösser et al., 2013). They demonstrate a successful attack using memory table exploitation. Moos, Wegener, and Moradi (2021) demonstrate the effectiveness of deep learning models when used during side-channel leakage assessment compared to more traditional statistical analysis-based side-channel leakage tests such as the Welch's t-test and the Pearson's χ^2-test. Moos et al. demonstrate the effectiveness of deep learning over traditional tests when the input trace data is misaligned, or the leakage model has a multivariate distribution. Wang et al. (2020) demonstrates the effectiveness of training a neural network on power trace side-channel data pulled from multiple homogenous and non-homogenous devices. They increased the success rate of attacks executed with fewer traces using a network trained on the homogenous and non-homogenous devices. Their research is based on the idea that signal and noise diversity can often help to solve problems more efficiently. Danial et al. (2022) also uses this principle, training a Deep Neural Network on EM traces captured from multiple devices to demonstrate the effectiveness of diversity when confronted with leakage that has a low signal-to-noise ratio (SNR). Das and Sen, (2020) uses test-vector leakage assessment and SNR as inputs into a greedy gradient-search heuristic to find a predicted optimal point to conduct leakage collection to minimize the time required to perform successful key retrieval.

2.2 Deep Learning

Deep learning is a form of machine learning that emphasizes learning successive layers of increasingly meaningful representations of data. Historical machine learning models relied on complex mathematical models and highly sanitized input to turn inputs and outputs into learned rules. The addition of many layers of neurons that add those meaningful representations of data allows deep learning to eliminate the need for those extra preprocessing steps. The ability of deep learning to forgo this excess work has contributed to the gain in popularity in recent years. Deep learning has been used in many fields and for many purposes, including image classification, speech recognition, natural language processing, autonomous driving, and strategy gaming. More recently, the field of side-channel research has found deep learning to be quite useful in performing SCA profiling attacks (Moos, Wegener, and Moradi, 2021)(Wang et al., 2020).

For our experiments, the input data $d_i = \mathbb{R}^e$ where $i \in \{1,2, \dots, N\}$ with N denoting the number of traces in our dataset, and e denoting the number of input features. Our input data is mapped to a set of labels L where $l(d_i) \in L$. And our labels are presented as a one-hot encoded vector $v = \{0,1\}^{|L|}$ where $m \in L$ and $v_m = \begin{cases} 1 \; if \; m = l(d_i) \\ 0 \quad otherwise \end{cases}$.

This research uses a convolutional neural network (CNN) which is a feed-forward deep learning algorithm. A convolutional neural network is made up of one input, several convolutional and hidden layers, dense layers, and one output layer. These layers often include pooling, batch normalization, and dropout layers. Pooling layers slide a two-dimensional filter over the feature map of the input to the layer. This filter summarizes the features within the region covered by the filter (Gholamalinezhad and Khosravi, 2020). Batch normalization serves to prevent overfitting of a model. It normalizes the output of each layer to enable the layers to perform learning independently. Dense layers are used as they are fully connected layers and facilitate the classification for the output layer. Each layer is made up of a matrix of neurons with initially randomized and later learned weights w that get applied to its inputs x linearly. The convolutions in a CNN generate features. Features are the measurable property found in the input data that allows the CNN to guess the correct output more accurately. Then a non-linear activation function is applied to each point in the resulting matrix. The output of a given layer is the input for the following layer. The final layer of a neural network is the output layer. In our case, we use a one-hot encoded vector to represent the classified output.

At the beginning of training all the weights are randomized. The input samples are separated into batches of a fixed size and all samples in a batch are evaluated simultaneously. Each time the network completes a training on a batch, the output layer presents a prediction y' for each of the inputs. These predictions are then compared to their corresponding label values and based upon the distance from the labels to the predictions, loss values are calculated using a loss function. Once all the losses are calculated, backpropagation is used to update the weights of all the neurons based on the gradient of the loss function. Each time covering the entire input dataset during training is called an epoch.

2.3 Side-Channel Analysis

Side-channel analysis is the analysis and exploitation of the weaknesses in the physical implementation of electronic devices. Often, SCA work is used against algorithms to recover the algorithm's secret information. This can be accomplished via observing power draw, EM emissions, and execution time of operations on a system. The purpose of SCA is to determine if there is discernable leakage from a given device, and if that leakage can lead to an attacker learning secret information from the device such as an encryption key. In the most effective SCA attacks, two phases are used. The first is a profiling phase in which the attacker uses a similar version of the device under attack to profile the device's leakage and create a model with which to execute an attack. The attack phase consists of gathering traces from the target device and running those traces through the profiling model to determine the key. During the attack phase, Bayesian analysis is used to execute a Maximum Likelihood strategy, which reduces to estimating the likelihood of each key byte candidate being the correct key byte and selecting the candidate with the highest probability. As each new trace is ran through the Bayesian process, the likelihood of each key candidate gets updated, and each candidate gets ranked in order of overall likelihood. The attack is considered a success if the correct key byte is ranked first in the ranked list (Benadjila et al., 2019).

2.4 Advanced Encryption Standard

The Advanced Encryption Standard is a widely used encryption algorithm. It uses three different key lengths, 128, 192, and 256 bits. For our experimental setup we use the 128-bit key length. The AES-128 algorithm uses 10 rounds of encryption, and each round is comprised of four steps: a non-linear substitution of bytes using the SBox lookup table, a row shift step, a column mixing step, and adding a round key. During the rounds of AES, the raw plaintext and raw key are only exposed in the first of the 10 rounds (Wang et al., 2020). Therefore, our experiment will focus on the trace data pulled from the first round of encryption. In the ChipWhisperer XMEGA's software implementation of AES, the rounds are completed in sequence. However, in many other AES implementation, including on FPGA devices, the rounds are completed simultaneously.

3. Experimental Design

This section describes the precise EM setup that we used to detect leakage during our collection. It further describes the neural network setup that we used to assess the leakage and carry out our attacks.

3.1 Collection

Our experiment sought to define a framework for determining the best EM collection point to perform EM SCA attacks on a given chip. Our setup consisted of the ChipWhisperer Lite XMEGA platform running a software implementation of the AES-128 encryption standard. Figure 1 displays the scanning setup. The ChipWhisperer device was scanned using a LeCroy WaveRunner 10xXi-A oscilloscope and a Riscure PR137LS probe with a diameter of 1.5mm. As the ChipWhisperer platform uses a software implementation of AES, side-channel leakage is not so highly localized that we would need a probe with a higher spatial sensitivity (Danial et al., 2022). As depicted in Figure 2, the scan was performed in a raster scan 5x5 grid. The XMEGA chip is 10mm by 10mm and we collected measurements every 2.5mm.

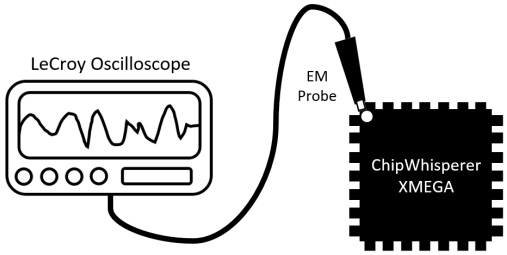


Figure 1: The ChipWhisperer is scanned using a near-field EM probe with data sent back to the LeCroy oscilloscope.

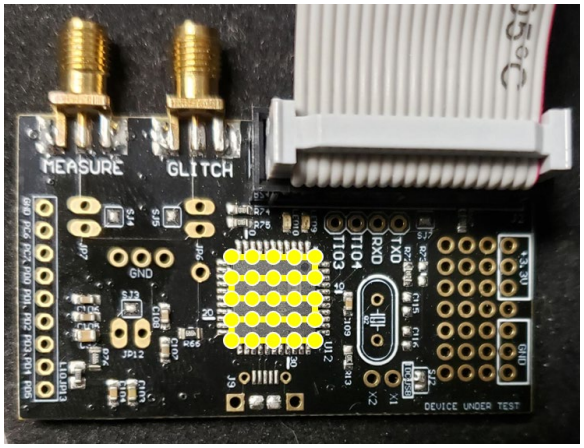


Figure 2: Depiction of the raster scan above the ChipWhisperer XMEGA.

The probe was moved across the board using an XYZ translation table. At each point, 24,000 profiling traces, 3,000 testing traces, and 3,000 attack traces were collected with a sampling rate of 100MHz and a sampling time of 5ms for a total of 500,000 samples per 500ms trace. Figure 3 shows the area we isolated which includes the first round of AES as this is the only time the real key is exposed in the SBox operation. This translates to the samples between 150,000 and 180,000. To collect our profiling traces, we generated randomized plaintext and randomized keys, and to collect our attack traces we generated randomized plaintexts and used a single consistent key across all 25 collection points.

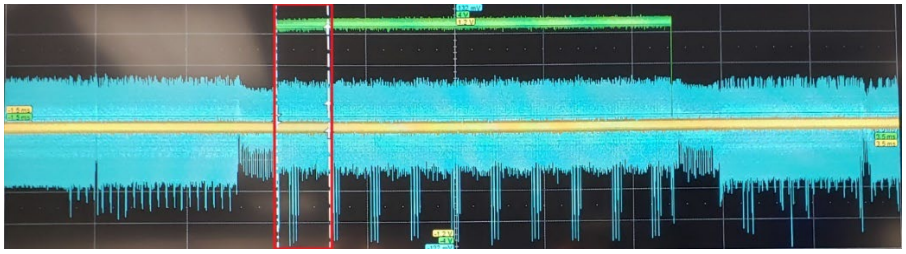


Figure 3: Oscilloscope output showing location of isolated samples

Prior to entering our data into our neural network, we create and attach the labels to our datasets. These labels are generated to match the output of the SBox operations. This simply means we assign the output of $p[1] \oplus k[1]$ to each trace (Benadjila et al., 2019). Then the collected profiling traces are separated out into the first 24,000 traces, used for training, and the final 3,000 traces, used for individual point testing. The training traces for each point are combined into one large training dataset consisting of 600,000 traces.

3.2 Neural Network

Our neural network, shown in Figure 4, consists of five convolutional blocks of increasing size. We used the following convolutional network with batch normalization, Rectified Linear Unit (ReLU) as our activation function, and a single pooling layer in each of our convolutional blocks.

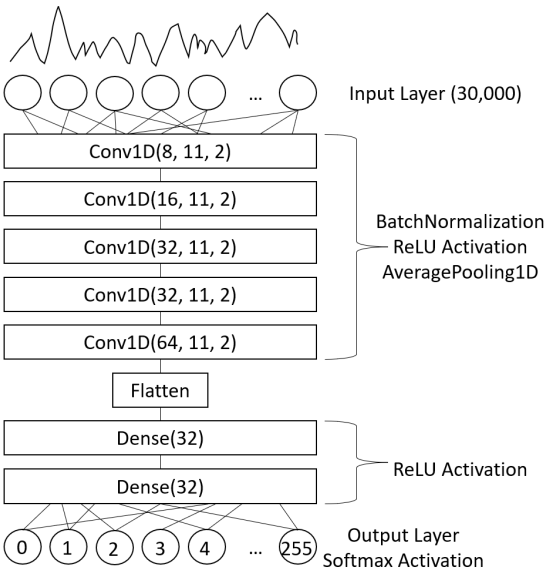


Figure 4: Architecture of the CNN. The network contains five convolutional blocks. Following each convolutional layer is a batch normalization layer, a ReLU activation function, and an average pooling layer. Two dense layers are used as the classification block, and the final layer uses softmax as the activation function and consists of a one-hot encoded vector that represents the predicted key byte output.

We used *CategoricalCrossentropy* as our loss function, *Adam* as the model optimizer, and *softmax* as the activation function of the final layer. Our network was built, tested an evaluated in the Python library Keras (keras-gpu version 2.9.0) using TensorFlow (tensowflow-gpu 2.10.0) as the backend. *Adam* is first introduced by Kingma and Ba, (2015) and sought to combine the benefits of *RMSPropagation* which performs well with on-line and non-stationary settings, and *AdaGrad* which performs well with sparse gradients (Tieleman, and Hinton, 2012)(Duchi, Hazan and Singer, 2011). We used a learning rate of 0.00002, a batch size of 1,000, and train over 50 epochs.

4. Experimental Results

This section describes the results provided by our framework using the above experimental setup. We trained 10 models using the architecture defined in Section 3.2. Then we tested the accuracy when running the model against the 3,000 attack traces collected at each of the 25 collection points.

4.1 Point Test Results

The average final training accuracy of the 10 trained models was 0.86%. Despite this, the overall average testing accuracy of all 10 models across all 25 points was 0.38%. This is almost equal to the likelihood of randomly guessing the correct key byte, or 1 in 256. Individual points did better than average chance and several points did worse. This means that the model was able to extract useful information from certain points that contributed to correctly guessing the key byte value. However, the lower accuracy points were less helpful in determining the key byte value and in some cases did worse than random guessing. Using the accuracies determined from testing all 10 models, we were able to define which points consistently resulted in higher test accuracy. Figure 5 displays the average accuracies of each collection point in a heat map format, and Figure 6 displays the 95% confidence intervals associated with each collection point. While we cannot conclude that point 6 is statistically better than all other points, points 6 and 10 are statistically significant when compared to one another. This means we can reject the null hypothesis that the location of the collection does not matter. Point 6 returned an average accuracy of .45% and point 10 returned an average accuracy of .33%.

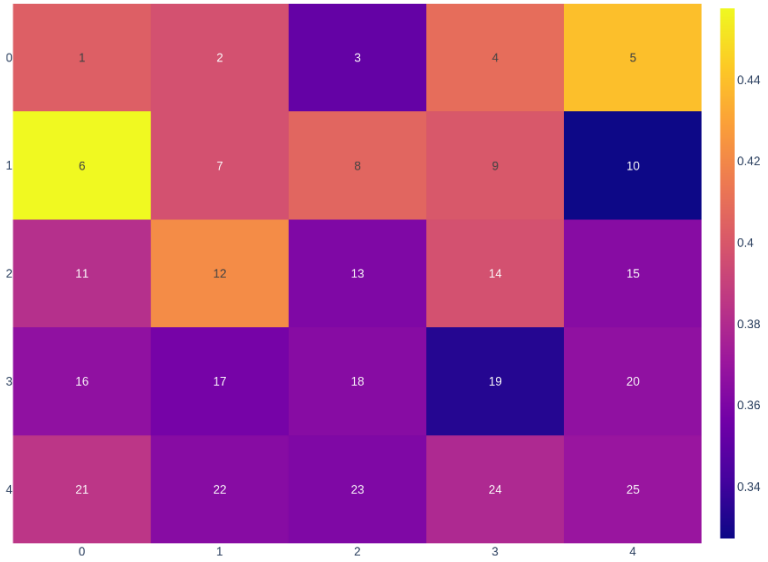


Figure 5: Heat map of average accuracies at each of the 25 points for all 10 trained models

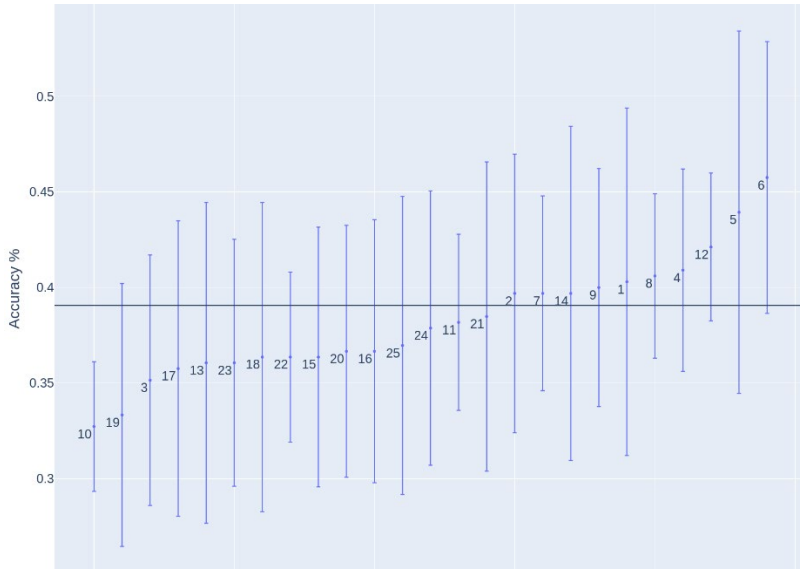


Figure 6: Confidence intervals for each of the 25 points. The horizontal line displays the likelihood of success with random guessing.

To further test the null hypothesis and determine if the accuracies returned by points 6 and 10 were statistically different, a right-tailed t-analysis is performed with an α of 0.05, 9 degrees of freedom, and means of 0.45% for point 6 and 0.33% for point 10. Using the t distribution table along with this alpha and degrees of freedom, results in a critical value of 1.833 (Montgomery, 2020). Using the right-tailed t-test results in a t value for this test of 2.828 which is higher than the critical value. Therefore, the null hypothesis, that position does not matter when collecting EM SCA leakage data, is rejected in favor of the alternate.

4.2 Attack Results:

To demonstrate the effectiveness of our framework, we must execute attacks at each of the identified points to show that the point with the higher test accuracy returns the key byte more effectively. We trained two models using the same CNN architecture defined in Section 3.2, one using the data at point 6 and one using the data at point 10. We then executed attacks using each model. We will refer to these models from here on as model 6 and model 10 respectively. Each attack randomized the order in which the attack traces were received by the Bayesian process to represent real attack scenarios more accurately. Each trace fed into the Bayesian process updates the maximum likelihood array of possible key byte values based upon the previous maximum likelihood array and the prediction returned by the model based upon the new trace.

The final training accuracy of the model from point 6 was 2.98%, and the accuracy at point 10 was 2.28%. The test accuracy at point 6 when tested using model 6 was 3.56%, and the test accuracy at point 10 when using model 10 was 2.4%. We then ran attacks against each model until we had 30 successful attacks. Model 6 was able to return 30 successes in 194 attacks, and model 10 was able to return 30 successes in 549 attacks. This translates to a success rate of 15.46% for model 6 and 5.46% for model 10. Model 6 was able to retrieve the key byte 2.83 times more often than model 10. This demonstrates the effectiveness of our framework in enabling more successful key recovery.

5. Conclusion

Our research demonstrates the success of our leakage search framework on the ChipWhisperer LITE XMEGA platform when tested against a software implementation of AES-128. We trained a CNN on EM data collected from several points over the device of interest and identified points that were statistically significantly different in their accuracy. We demonstrated that training a CNN on the leakage data collected at the identified points with statistically higher accuracy and lower accuracy in a distinguishable difference in the likelihood for success when conducting a side-channel attack. However, the ChipWhisperer Lite platform was developed as a teaching mechanism for those new to the world of SCA, and as such this same research and framework must be demonstrated as successful on other more secure platforms and against common countermeasures such as introducing random noise and lower-level metal routing. This framework could also be tested for effectiveness against a device implementing a hardware-based AES scheme. Other researchers have also used other algorithms to determine the point on a chip with the highest perceived leakage. Further research should be done to understand the effectiveness of both approaches compared to one another to determine the better approach.

References

Andrikos, C., Batina, L., Chmielewski, L., Lerman, L., Mavroudis, V., Papagiannopoulos, K., Perin, G., Rassias, G. and Sonnino, A. (2019). Location, Location, Location: Revisiting Modeling and Exploitation for Location-Based Side Channel Leakages. Lecture Notes in Computer Science, pp.285–314. doi:10.1007/978-3-030-34618-8_10.

Benadjila, R., Prouff, E., Strullu, R., Cagli, E. and Dumas, C. (2019). Deep learning for side-channel analysis and introduction to ASCAD database. Journal of Cryptographic Engineering, 10(2), pp.163–188. doi:10.1007/s13389-019-00220-8.

Danial, J., Das, D., Golder, A., Ghosh, S., Raychowdhury, A. and Sen, S. (2022). EM-X-DL: Efficient Cross-device Deep Learning Side-channel Attack With Noisy EM Signatures. ACM Journal on Emerging Technologies in Computing Systems, 18(1), pp.1–17. doi:10.1145/3465380.

Das, D. and Sen, S. (2020). Electromagnetic and Power Side-Channel Analysis: Advanced Attacks and Low-Overhead Generic Countermeasures through White-Box Approach. Cryptography, 4(4), p.30. doi:10.3390/cryptography4040030.

Duchi, J., Hazan, E. and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. The Journal of Machine Learning Research, 12.

Gholamalinezhad, H. and Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, a Review. Computer Vision and Pattern Recognition.

Kingma, D. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. International Conference for Learning Representations.

Masure, L., Dumas, C. and Prouff, E. (2019). A Comprehensive Study of Deep Learning for Side-Channel Analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems, pp.348–375. doi:10.46586/tches.v2020.i1.348-375.

Montgomery, D.C. (2020). Design and analysis of experiments. Hoboken, Nj: Wiley.

Moos, T., Wegener, F. and Moradi, A. (2021). DL-LA: Deep Learning Leakage Assessment. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp.552–598. doi:10.46586/tches.v2021.i3.552-598.

Schlösser, A., Nedospasov, D., Krämer, J., Orlic, S. and Seifert, J.-P. (2013). Simple photonic emission analysis of AES. Journal of Cryptographic Engineering, 3(1), pp.3–15. doi:10.1007/s13389-013-0053-7.

Tieleman, T. and Hinton, G. (2012) Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA: Neural Networks for Machine Learning, 4, 26-31.

Wang, H., Forsmark, S., Brisfors, M. and Dubrova, E. (2020). Multi-Source Training Deep-Learning Side-Channel Attacks. 2020 IEEE 50th International Symposium on Multiple-Valued Logic (ISMVL). doi:10.1109/ismvl49045.2020.00-29.

Using Deep Reinforcement Learning for Assessing the Consequences of Cyber Mitigation Techniques on Industrial Control Systems

Terry R. Merz and Romarie Morales Rosado
State University of New York, College of Homeland Security and Cybersecurity
Pacific Northwest National Laboratory

tmerz@albany.edu
romarie.morales@pnnl.gov

Abstract: This paper discusses an in-progress study involving the use of deep reinforcement learning (DRL) to mitigate the effects of an advanced cyber-attack against industrial control systems (ICS). The research is a qualitative, exploratory study which emerged as a gap during the execution of two rapid prototyping studies. During these studies, cyber defensive procedures, known as "*Mitigation,* were characterized as actions taken to minimize the impact of ongoing advanced cyber-attacks against an ICS while enabling primary operations to continue. To execute *Mitigation* procedures, affected ICS components required rapid isolation and quarantining from "healthy" system segments. However today, with most attacks leveraging automation, mitigation also requires rapid decision-making capabilities operating at the speed of automation yet with human-like refinement. The authors settled on the choice of DRL as a viable solution to this problem due to the algorithm's designs which involves "intelligent" decisions based upon continuous learning achieved through a rewards system. The primary theory of this study posits that processes informed by data sources relative to the execution path of an advanced cyber-attack as well as the consequences of deploying a particular *Mitigation* procedure evolve the system into an ever-improving defensive capability. This study seeks to produce a defensive DLR based software agent trained by a DRL based offensive software agent that generates policy refinements based upon extrapolations from a corrupted network state as reported by an IDS and baseline data. Results include an estimation rule that would quantify impacts of various mitigation actions while protecting the operational critical path and isolating an in-progress attack. This study is in a conceptual phase and development has not started. This research questions for this study are: RQ1: Can this software agent categorize correctly an in-progress cyber-attack and extrapolate the potential ICS assets affected? RQ2: Can this software agent categorize novel cyber-attacks and extrapolate a probable attack vector while enumerating affected assets? RQ3: Can this software agent characterize how operations are affected by quarantine actions? RQ4: Can this software agent generate a set of ranked recommended courses of action by effectiveness, and least negative effects on the operational critical path?

Keywords*:* Deep reinforcement learning, APT, ICS, Mitigation, AI based automated attacks, ICS cyber-attacks, energy security, cybersecurity

1. Introduction

The conceptual framework for this study emerged as a research gap during the execution of two Department of Defense rapid prototyping studies called JBASICS (2017) and MOSAICS (2021). The first study, JBASICS involved the development of written tactics, techniques, and procedures (TTP) that ICS operators could use to detect, mitigate, and recover from an advanced cyber-attack (J-BASICS 2017). The second study, MOSAICS sought to automate a set defined during the JBASICS study (DoD JCTD 2021). In both studies, *Mitigation* was characterized as actions taken to minimize the negative effects of an on-going advanced cyber-attack against an industrial control system (ICS) while simultaneously enabling the organization to maintain primary operations. However, *Mitigation* procedures were only lightly treated during the execution of the two studies as baselining and detection technologies for ICS were evolving and had not reached a maturity level that would allow the implementation of *Mitigation techniques* as defined by the programs (DoD MOSAICS JCTD 2021).

Lastly, to effectively execute a *Mitigation* procedure, ICS components compromised by an advanced cyber-attack would require rapid isolation and quarantining from "healthy" system segments. However, this would need a decision-making capability that operates at the speed of an automated decision-making process, yet has a human-like refinement, and nuanced understanding around such decisions (DoD MOSAICS JCTD 2021).

2. Research Design

As described by the CISA Alert AA22-103A (2022) adversarial tactics are continuously evolving. Therefore, the Merz/Morales research team hypothesizes that success of a *Mitigation* technology rests on adaptive processes. Additionally, the research team deducted this requirement precludes the use of rule-based algorithms such as

Decision Trees, or even potentially traditional Machine Learning, which requires training sets. However, both deep learning and reinforcement learning represent forms of autonomous learning that apply continuous learning (Nguyen, Reddi 2019).

With deep learning requiring pattern inputs to train the algorithm, reinforcement training offers a solution using trial and error with rewards (Nguyen, Reddi 2019). Both approaches have benefits and drawbacks. However, the research team hypothesized that when used in combination, i.e., Deep Reinforcement Learning (DRL) could potentially achieve the decision-making refinement described and required in the *Mitigation* conceptual framework used in both the J-BASICS and MOSAICS prototyping efforts (J-BASICS 2017).

DRL has demonstrated rapid decision-making capabilities in areas such as robotics, autonomous surgery, autonomous vehicles, biological data mining and drug design (Nguyen, Reddi 2019). Recent developments in cybersecurity research relative to the Internet of Things (IoT) have been centered around the use of DRL and have produced noteworthy results such as a DRL based resource allocation framework for "Smart Cities" that successfully integrates processes from networking, caching, and computing (Ferdowsi, et al., 2018). To that end, this study seeks to extend existing research using DRL in cybersecurity and to produce a software agent that generates policy refinements based upon extrapolations from a corrupted network state as reported by an IDS and baseline data. These corrupted states would have been produced by an offensive software agent also using DRL. The defensive software agent will include an estimation rule that quantifies the impacts of the various possible *Mitigation* actions on protecting the operational critical path while isolating the in-progress attack.

Success criteria: While the success of this exploratory study is framed within the context of the 4 research questions, the possibilities exist that an offensive software agent training a defensive software agent will generate previously undefined outcomes. Therefore, the success of this study is not limited to the research questions being answered in the affirmative but includes a collection of additional system behaviors that will be evaluated from a qualitative perspective.

Validation: After training the computational model for proper classification we will move to model validation. The computational model is verified against the first set of real data and re- verified against a second or subset of real data.

3. Research Team:

The research team involves a partnership between the College of Emergency Preparedness, Homeland Security and Cybersecurity (CEHC) at the University at Albany and the Pacific Northwest National Laboratory (PNNL). These organizations bring together a cross disciplinary team consisting of cybersecurity subject matter experts relative to industrial control systems, and subject matter experts in Machine Learning.

4. Limitations of the Study:

This study is limited to industrial control systems and more specifically to electrical systems. The limitation is driven by the ability to baseline and model system behaviors that have regular, predictable traffic patterns.

5. Research Plan:

While the research team has determined that DRL in general is the most logical approach to solving the *Mitigation* problem, there are many algorithms that can be used to implement it. The two approaches selected for this study are:

Table 1: Deep Reinforcement algorithm candidates (Nguyen, Reddi 2019)

Proposed Agent Name	Application Focus Area	Algorithm	Model	Actions	Rewards
Mitigation-TRPO	Increasingly improved defensive capabilities against cyber actors: devises filtering schemas to detect corrupted states and respond to adversarial actions.	Trust Region Policy Optimization (TRPO)	Model-Free	Determines which extrapolation rule to apply to generate an estimated viable operational state from a corrupted state.	Function: using state inputs
Mitigation-DDQN+A3C	Autonomous defense to include in Software	Double Deep Q Network and Asynchronous	Model-Free	Leverages dueling agents: attack agent learns to select most viable attack vector	Function: rewards based upon

Proposed Agent Name	Application Focus Area	Algorithm	Model	Actions	Rewards
	Defined Networks (SDN)	Advantage Actor-Critic Algorithm		while defending agent can take four actions: isolate, patch, reconnect or migrate to protect an asset and preserve as many nodes along the operational critical path.	the status of the assets along the critical path.

With the Mitigation DDQN+AC3 approach likely to produce the desired agent, the Mitigation-TRPO process would allow a Human-in-the-Loop to make an informed decision about proposed Mitigation actions. Therefore, we should explore using both to achieve different objectives of the Mitigation process.

6. Research Phases with Expected Outcomes:

The conceptual framework for this study is centered around the possibility that an industrial control system can be segmented in such as manner as to isolate an incoming cyber-attack while maintaining primary systems operations.

Using the Perdue reference model as the basis for the study, it is assumed the defensive software agent will focus on segmenting network traffic by one of the following methods (or a combination thereof) by user accounts, processes, applications, and protocols. It is also expected the defensive software agent may identify a physical action to be the most reasonable and will thus alert the operator to physically disconnect a component of the network.

Conversely, the offensive software agent, will mimic the behaviors of actual automated attacks having been trained in cyber-attack patterns typical of an advanced persistent threat. It will identify user accounts, processes, applications, and protocols for exploitation with the objective of terminating primary systems operations. If either software agent (offensive or defensive) are successful, a point award is applied to the software agent, and the agent "learns" the methodology used was successful and the method is captured as a sample of desired actions.

The inclusion of a reward system in training the respective software agents differs from traditional Deep Learning in that the software agents learn on a continuous basis. This approach represents a next level of intrusion detection that not only identifies activities that could lead to a corrupted network state (behavior observation), but also extrapolates courses of actions operators could immediately take to defend an industrial control system and maintain primary operations.

Figure 1 illustrates the communications process by which the defensive and offensive software agents interact with the industrial control system environment.

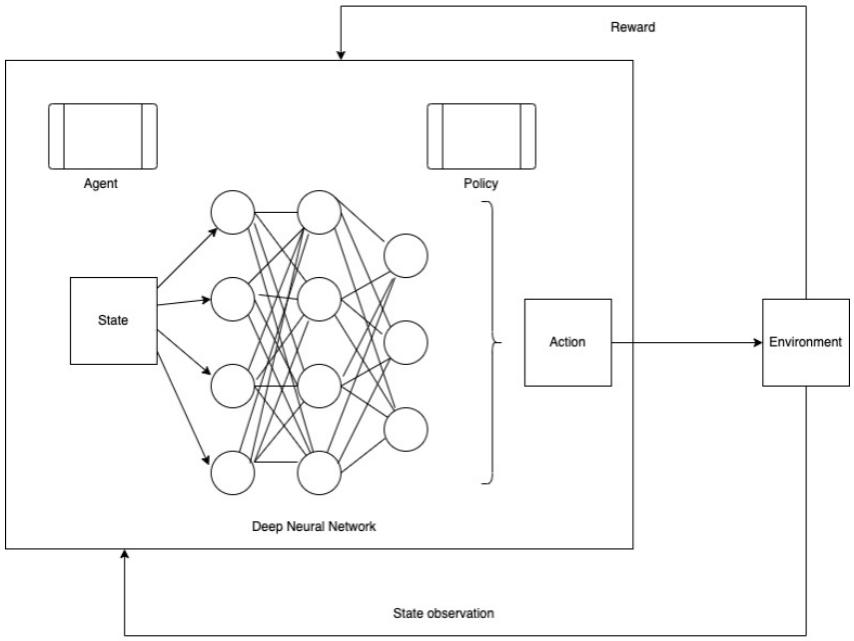


Figure 1: Conceptual Framework (Merz/Morales 2022) - Offensive/Defensive DRL Operations

7. In Conclusion:

Once funding is identified and secured, this study will produce 2 software agents using deep reinforcement learning: an offensive and a defensive agent. The project's success trajectory will be evaluated on a continuous basis throughout the execution of the study via a project risk management plan. Quarterly reviews will be conducted to determine the progress of the study's critical path, potential blockers, and mitigations to blockers. Since the project will be using an Agile development process, regular "hot washes" will be conducted at the conclusion of each sprint and procedures adjusted as needed. The strategic success of the project will be evaluated by the number and quality of peer-reviewed journal papers, follow on research, resultant thesis and dissertations, and acceptance into a transition to practice study resulting in the final fielding of the defensive software agent.

The project success metrics will be included in final reports, which will also address unexpected events, successes, failures, completeness of assumptions, and insights into future applications of the defensive agent.

The resultant models, artifacts and findings from this project are designed to first and foremost produce the foundations of a mitigation capability (in the form of a defensive software agent). This will position the defensive software agent for a transition to practice. A potential target framework for this transition is the Integrated Adaptative Cyber Defense framework, or IACD. The development of this framework was sponsored by the National Security Agency and developed by Johns Hopkins Applied Physics Laboratory (JHU APL N/D).

References

Cybersecurity & Infrastructure Security Agency (2022) "APT Cyber Tools Targeting ICS/SCADA Devices", Alert (AA22-103A), U.S. Department of Homeland Security, Washington D.C.

Department of Defense (DoD), Joint Technical Demonstration (JCTD) (2021) "More Situational Awareness for Industrial Control systems (MOSAICS)", Washington D.C.

Ferdowsi, A., Challita, U., Saad, W. and Mandayam, N.B., "Robust deep reinforcement learning for security and safety in autonomous vehicle systems," in 21st International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 307-312.

Johns Hopkins Applied Physics Laboratory (JHU APL), (N/D) "About IACD", https://www.iacdautomate.org/aboutiacd, Laurel, Maryland.

Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nahvandi, S. (2019), "Deep Learning for deepfakes creation and etection: a survey" arXiv preprint arXiv:1909.11573.

Nguyen, T.T., Reddi, V.J. (2019) "Deep reinforcement Learning for Cybersecurity". IEEE Transaction on Neural networks and Learning Systems, 2021. https://doi.org/10.1109/TNNLS.2021.3121870.

United States Cyber Command, Department of Defense (2017) "Joint Base Architecture for Security Industrial Control Systems" (J-BASICS), Fort Meade, Maryland.