Sabermetrics for Cyber: Collecting and Analyzing User Activity Data from Ephemeral Exercises

Jael Rivera, Jarrett Booz and Josh Hammerstein
Software Engineering Institute, Carnegie Mellon University, Pittsburgh, USA

jrivera@sei.cmu.edu
jabooz@sei.cmu.edu
joshh@sei.cmu.edu

Abstract: The term *sabermetrics* was coined in the 1970s by members of the Society for American Baseball Research (SABR) to describe how baseball teams use advanced analytics to evaluate talent and maximize performance both offensively and defensively. Sabermetrics transformed professional baseball through its data-driven approach, enabling teams to devise new tactics and strategies for improving individual and overall team performance. The concept of sabermetrics or advanced analytics can also be applied to the cybersecurity domain to improve performance, both offensively and defensively, and to better evaluate talent. To do this, data is needed. Cybersecurity exercises are well suited for providing this data because they are designed to develop critical technical skills in controlled, simulated environments that closely mirror real-world threats. However, preserving data for ephemeral cybersecurity exercises can be challenging because these environments are temporary, and when they are torn down, log data is lost unless deliberate actions are taken to retain the data for future use. This includes all information regarding the actions participants took in the exercise. Recognizing that important information can be gleaned by analyzing this data, the Software Engineering Institute (SEI) at Carnegie Mellon University developed a capability to capture a high-fidelity record of user activities during cybersecurity exercises. This paper discusses the motivation behind this development, the insights that can be gained from the collected data, and how the SEI configures exercises used in cybersecurity competitions to collect and store user activity data for future detailed analysis.

Keywords: Cybersecurity exercises, Cyber workforce development, Cybersecurity training, Data collection, Data analysis, Performance analytics

1. Introduction

The term *sabermetrics* was coined in the 1970s by members of the Society for American Baseball Research (SABR) to describe how baseball teams use advanced analytics to evaluate talent and maximize performance both offensively and defensively (Wikipedia, n.d.) (Lewis, 2004). Sabermetrics transformed baseball through its data-driven approach, enabling teams to devise new tactics and strategies for improving individual and overall team performance. For example, by analyzing data on where individual batters tended to hit the ball, teams could place fielders where they would be more likely to receive those hits, often in obscure and nontraditional positions. These defensive shifts created competitive advantages so effectively that Major League Baseball eventually banned them.

The concept of sabermetrics or advanced analytics can also be applied to the cybersecurity domain to improve performance, both offensively and defensively, and to better evaluate talent. By creating exercise environments that collect data about users' actions and performance, we allow for analysis of that data, unlocking the ability to (1) identify and devise effective cybersecurity tactics and strategies and (2) develop predictive models and data-driven evaluations for cybersecurity talent.

2. Background

2.1 The President's Cup Cybersecurity Competition

The President's Cup Cybersecurity Competition was established by Executive Order 13870 as a federal initiative aimed at identifying, developing, and recognizing top cybersecurity talent among U.S. government employees (Executive Office of the President, 2019) (Cybersecurity and Infrastructure Security Agency (CISA), n.d.). Created in response to the growing need for skilled cybersecurity professionals within the United States government federal workforce, the competition is designed to foster a culture of continuous learning and improvement in cybersecurity skills. Federal employees from various agencies participate in this competition, which not only challenges their knowledge and skills but also provides them with an opportunity to apply their expertise in a competitive and collaborative environment.

The competition leverages a variety of cybersecurity exercises that are tailored to test the participants' abilities across a wide range of cybersecurity domains. These exercises are based on real-world scenarios, making the competition an effective tool for both assessment and professional development. To efficiently

deliver these challenges, the competition uses ephemeral cybersecurity exercises, which offer a dynamic and flexible platform for participants to engage with the material.

The President's Cup Cybersecurity Competition is run by the Department of Homeland Security (DHS) Cybersecurity and Infrastructure Security Agency (CISA). The Carnegie Mellon University (CMU) Software Engineering Institute (SEI) helped CISA facilitate the competition by building the infrastructure and exercises used within it from the inaugural competition through President's Cup V.

2.2 Ephemeral on-demand Cybersecurity Exercises and Their Uses

Cybersecurity professionals are constantly seeking innovative ways to enhance their expertise. One way to do this is through cybersecurity exercises that provide users with realistic environments to hone and develop cybersecurity skills.

Cybersecurity exercises typically instantiate one or more virtual machines (VMs) configured to simulate enterprise systems and networks. These are delivered on demand so that users can deploy the exercise and associated VM resources whenever time permits. On-demand exercises are built to be ephemeral, meaning all resources used during an exercise session are destroyed once it is complete, reducing costs and freeing up computing resources for the provider.

Cybersecurity exercises are often used for both assessment and training of cybersecurity practitioners. When using cybersecurity exercises as an assessment, administrators use completion criteria to determine if the participants have successfully completed the exercise. When using cybersecurity exercises as training, participants are given instructions with tasks to complete and steps to follow to complete those tasks. The level of detail in the instructions can be adjusted depending on the skill level of the participants. The technical report Challenge Development Guidelines for Cybersecurity Competitions provides more details on how cybersecurity exercises can be designed and developed for use in a competition format to assess cybersecurity skills (Booz, et al., 2022).

2.3 Cybersecurity Work Roles and Tasks

The National Institute for Standards and Technology (NIST) published the Workforce Framework for Cybersecurity (NICE Framework) to standardize the naming of cybersecurity work roles (i.e., job functions) and associated tasks, knowledge, and skills (National Initiative for Cybersecurity Careers and Studies, n.d.). As we build cybersecurity exercises, we map exercise objectives to NICE Framework work roles and tasks. This mapping provides exercise administrators and participants insight into specific cybersecurity topics and skills the exercise addresses. If the exercise is used for assessing cybersecurity practitioners, administrators can point to specific tasks or skills that participants are strong in or need to improvement on. If the exercise is used for training, administrators can organize content based on work roles or tasks, and participants can easily identify exercises that target the skills they are looking to develop.

2.4 The need for Enhanced Logging in Ephemeral Exercises

While hosting the President's Cup Cybersecurity Competition, we encountered a significant challenge related to the ephemeral nature of the exercises. Since these environments were designed to be temporary, all data was destroyed once the exercise session ended, including log data. Without access to log data, we had a hard time troubleshooting issues users encountered during the exercise. Specifically, we could not retrace their steps in the exercise and had to rely on their descriptions of the problem, which were often incomplete. This approach was time consuming and did not always yield accurate results.

Recognizing these limitations, we saw the need to develop an enhanced logging mechanism to extract and collect logs from these ephemeral exercises. By preserving detailed logs of all activities that occurred within the exercise environments, we could troubleshoot more effectively and ensure that the exercises were deploying correctly and working as intended. This shift to enhanced logging marked a significant improvement in our ability to manage the competition. Beyond simply aiding in troubleshooting, the logs provided us with a new level of understanding regarding participant actions and behaviors during the exercises.

2.5 Leveraging Data for Cybersecurity Insights

As we continued to collect data from these ephemeral exercises, it became clear that the information at our disposal could be used for more than just troubleshooting. Since the logs provide a detailed record of each participant's actions during an exercise, we were able to use them to see exactly how participants interacted with our challenges, including the steps they took and the tools they used to navigate the exercises. We also

realized we could use this data to explore broader patterns and trends. This data collection has therefore become an invaluable asset not only for operational support but also as a foundation for more comprehensive analysis.

3. User Activity Data

While participating in cybersecurity exercises, users generate data about their activity that can provide valuable insights about their knowledge and skill. This data includes things such as command history and applications used during the exercise. In this section, we describe the data that is being collected and the infrastructure used to collect the data.

3.1 User command History and Graphical Application use

While participating in cybersecurity exercises, users enter commands on the exercise systems to run tools, perform analysis, and complete exercise tasks. Most commands are entered into a terminal application (command prompt). We capture the commands entered by the users and the output that each command generates, allowing us to reconstruct what each user did and saw through the terminal. This data allows exercise administrators to follow the command line actions taken by each user, step by step, and opens opportunities for more detailed analysis. However, not all actions occur in the command terminal. Many tools used during exercises involve graphical applications. A web browser is a simple example of a graphical tool, but other examples include Wireshark for packet analysis (Wireshark, n.d.), Burp Suite for website security testing (PortSwigger, n.d.), Autopsy for examining forensic images (Sleuth Kit Labs, n.d.), and Visual Studio Code for development (Microsoft, n.d.). Capturing graphical application usage alongside command history ensures we have a full picture of all tools used and actions taken by users during an exercise.

3.2 Log collection Infrastructure

To support the export and storage of desired log data from ephemeral, on-demand exercises, we configure the log collection infrastructure before exercises begin. This infrastructure connects two distinct networks:

- Ephemeral Network: A temporary network hosting the exercise environment, where main VMs are deployed for users.
- Persistent Network: A permanent network used for storing and analyzing logs, where the Graylog VM resides.

This infrastructure consists of three components (depicted in Figure 1):

- Exercise Environment (yellow, star): Includes one or more main VM(s), which are instantiated on the ephemeral network and provided to the user upon request so that they can complete the exercise.
- Administrative Server (green, diamond): The system used to configure and grade an exercise. This VM is connected to both the ephemeral and persistent networks and serves as a link for transferring the exercise data to the log collection environment.
- Log Collection Environment (blue, triangle): Designed to aggregate and store exercise logs permanently using a persistent network that uses a Graylog VM running Graylog log management software (Graylog, n.d.).

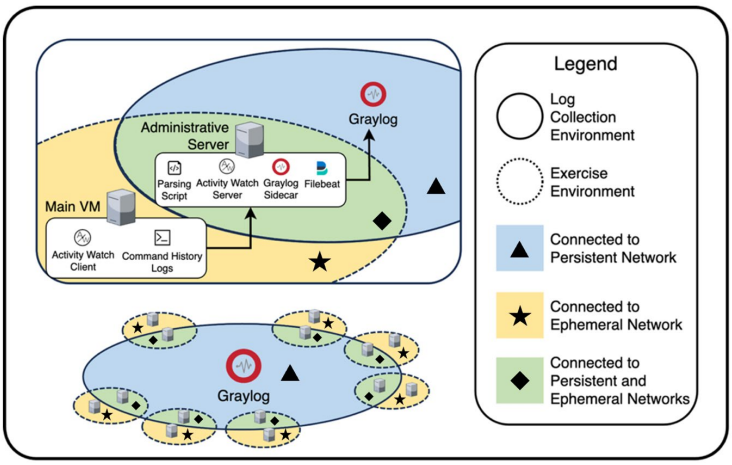

Figure 1: Exercise Log Collection Infrastructure

During an exercise, data flows from the main exercise VM (yellow, star) to the exercise administrative server VM (green, diamond), which then forwards the data to the Graylog VM (blue, triangle). This setup allows the multiple ephemeral exercise environments to be created and managed while maintaining a consistent connection to a single, persistent log collection environment. Once the exercise concludes, the exercise environment and the administrative server VMs are torn down, but the data remains stored in the Graylog VM.

3.2.1 Configuring the main VM(s) for logging

Users interact with the main VMs through both terminal commands and graphical applications. We implemented mechanisms to capture this activity and transmit it to the administrative server.

Command history

We developed a solution that captures command history and output in real time using a pair of custom scripts in conjunction with built-in tools such as the script and tail commands (Man7.org, n.d.). Each log entry includes the hostname, username, working directory, and timestamp, enabling the reconstruction of user actions. The logs are forwarded to the administrative server using rsyslog for processing and eventual export.

Graphical applications

To automatically track and monitor how time is spent on different applications, we use an open source application called ActivityWatch (ActivityWatch, n.d.). The main VM acts as an ActivityWatch client, recording application usage and sending logs to the ActivityWatch server running on the administrative server VM.

3.2.2 Configuring the administrative server for logging

The administrative server is the exercise's hub for log aggregation, bridging the ephemeral exercise network and the persistent logging network. Logs generated by the main VMs (as described in Section 3.2.1) are transmitted to the administrative server, where they are cleaned, processed, and formatted before being shipped to the Graylog server for permanent storage.

We use the administrative server for two primary reasons: to ensure logs are properly formatted before ingestion into Graylog and to prevent users from accessing others' log data or solutions. The server is not accessible to users and is the only system permitted to connect both networks, segmenting the infrastructure to maintain fairness and security during exercises and competitions.

Log Processing

While the ActivityWatch server exports events in JSON format (json.org, n.d.), its multiline structure is not easily parsed by Graylog. We developed a script to export and parse these JSON events, extract the system's hostname, application title, and duration of use, and transform them into a one-line format accepted by the Graylog server.

Log Transmission

To forward logs to Graylog, we use Filebeat, a lightweight log shipper (Elastic, n.d.). Configuration is centrally managed using Graylog's Sidecar feature (Graylog, Inc., n.d.), ensuring consistency across all administrative servers.

3.2.3 Configuring the Graylog server for log storage

To store the collected exercise logs permanently, we use Graylog, an open source, centralized log management solution with capabilities to capture, store, and conduct real-time analysis of data. In our infrastructure, a Graylog server is installed on a VM connected to a persistent network separate from the exercise environment (shown in the blue sections, marked with triangles, of Figure 1). This allows exercise administrators to troubleshoot, analyze and visualize exercise data (see Figure 2 and Figure 3 for examples of Graylog-generated data visualizations).

3.3 Exercise log Isolation

When facilitating exercises, there is a many-to-many relationship between exercises and users. This means:

- Many different exercises can be available to play at one time.
- A single user can deploy more than one exercise simultaneously.

- Many users can have exercises deployed simultaneously.

As an example, our system has more than 200 exercises available to play and more than 5,000 users, each of whom can deploy multiple exercises at any time. Since Graylog is a centralized logging server that stores logs from all exercises, each exercise is assigned a unique identifier (e.g., c11), and every deployment of an exercise receives a randomly generated isolation tag. For example, if Player 1 and Player 2 deploy exercise c11, their deployments might be tagged as **abcd1234** and **a1b2c3d4**, respectively. These identifiers allow administrators to query logs by exercise and user, ensuring scalability and accurate analysis. Because of the unique exercise identifiers and isolation tags, we can have thousands of exercises simultaneously deployed and forwarding their logs for persistent storage on Graylog.

4. Analysis of User Activity Data

Following the users' actions based on collected command history and application usage, we can extrapolate an exercise solution strategy for each participant. We can also compare the solution strategies employed by different participants. This can lead to asking and answering questions like "How many participants of this exercise follow the same solution path and use the same tools?" By parsing out and comparing strategies, we can evaluate their relative effectiveness and gain understanding of how different levels of expertise influence decision-making processes.

Insights gained from these comparisons help us visualize which sets of commands and tools were most instrumental in finding solutions during the exercises. Tying these commands and tools back to the NICE work roles mapped to the exercise can help exercise administrators draw conclusions like "Experts in work role X should show proficiency in using tools Y and Z."

Additionally, following a user's actions in the exercise from beginning to end can highlight a participant's ability to adapt to unexpected challenges, their proficiency with specific command line tools, and their ability to optimize processes under pressure. We can see which tools and commands a participant tried first. Using the output and the participant's subsequent steps, we can extrapolate if the tool produced the expected result, leading to the participant progressing in the exercise, or if the participant had to adapt their strategy to account for an unexpected outcome.

4.1 Analysis of Data for a Single Participant

Table 1 shows a selection of one participant's command history as they stepped through an exercise.

- In Index 1, we see the user performing some information gathering with nmap, a port scanning tool (nmap.org, n.d.).
- In Index 2 through Index 4, the user found that the target machine was a website with a database connected to it and decided to proceed with using the sqlmap tool (sqlmap.org, n.d.), trying several variations of sqlmap commands and using the tool's help menu along the way.
- In Index 5, the user then used an interactive SQL shell to further exploit a discovered SQL injection vulnerability.

This demonstrates an example of an experienced user following a process to gather information and iteratively solve a problem while using the help menus and tools they have at their disposal.

Table 1: Command history example

Index	Timestamp	Command
1	2024-04-16 10:37:31	nmap -Pn -sT -n 10.2.2.100 -p-
2	2024-04-16 10:38:51	sqlmap -u http://reactors.merch.codes --help
3	2024-04-16 10:41:12	sqlmap -u http://reactors.merch.codes/Home/Logs --help | grep cookie
4	2024-04-16 10:43:32	sqlmap -u http://reactors.merch.codes/Home/Logs --cookie="loggedin=true" --data='query="asdf"' --level=5 --help
5	2024-04-16 10:53:44	sqlmap -u http://reactors.merch.codes/Home/Logs --cookie="loggedin=true" --data='query="asdf*"' --level=5 --sql-shell

4.2 Comparing Data from Participants who did and did not Solve an Exercise

We can also compare the actions of users who solved an exercise to those who did not solve the exercise. Figure 2 highlights the differences in application usage between two users who attempted the same challenge

that was tagged with the Digital Forensics NICE Work Role (National Initiative for Cybersecurity Careers and Studies, n.d.).

- Player 1 successfully solved the exercise by completing all tasks and spent roughly 36 minutes (2,200 seconds) completing the exercise:
- *Autopsy: 19 minutes*
- *command terminal: 15 minutes*
- *VS Code: 2 minutes*
- Player 2 did not successfully complete the exercise despite spending 65 minutes on their attempt:
- *Autopsy: 56 minutes*
- *command terminal: 9 minutes*

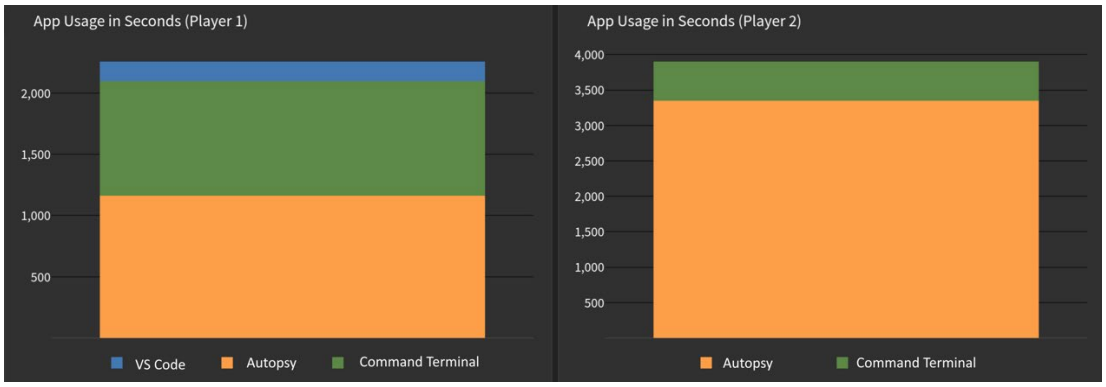

Figure 2: Application Usage: Solved vs. Did Not Solve

These two players used the same tools, though one was successful in solving the exercise and one was not. Because Player 1 successfully solved the challenge in 36 minutes, while Player 2 was unable to solve the challenge in 65 minutes, we can conclude that Player 1 has greater mastery of the Autopsy tool than Player 2. Player 1 may also be a better candidate than Player 2 to complete other tasks in the Digital Forensics NICE Work Role.

Furthermore, we can extrapolate beyond two users to all of the users who attempted an exercise as well. For example, if all users who attempted an exercise used the "Autopsy" tool for a majority of the time in the exercise, we can conclude that Autopsy is important to the tasks this exercise assesses under the Digital Forensics NICE Work Role. If users successfully completed the exercise, they demonstrated knowledge and experience using the Autopsy tool. If users did not successfully complete the exercise, they may be candidates for further digital forensics training using Autopsy.

4.3 Analyzing Data from Teams of Users

We can also use the collected user activity data to determine how teams of users are separating duties. In Figure 3, we show the activity data for four members of a team participating in the same exercise. Notice that Player 1 used many different tools (command terminal, Firefox web browser, Ghidra reverse engineering software, Wireshark packet analyzer, etc.). Player 2 used fewer tools than Player 1 (Firefox, Wireshark, command terminal, etc.). Player 4 only used one tool, and Player 3 did not interact with their VM at all.

From this, we can see that Player 1 was likely the leader of this team or had the most experience with what the exercise was targeting. Player 2 was involved with solving the exercise as well. Players 3 and 4 were most likely providing verbal input on how to solve the exercise without much or any interaction with the exercise VMs.

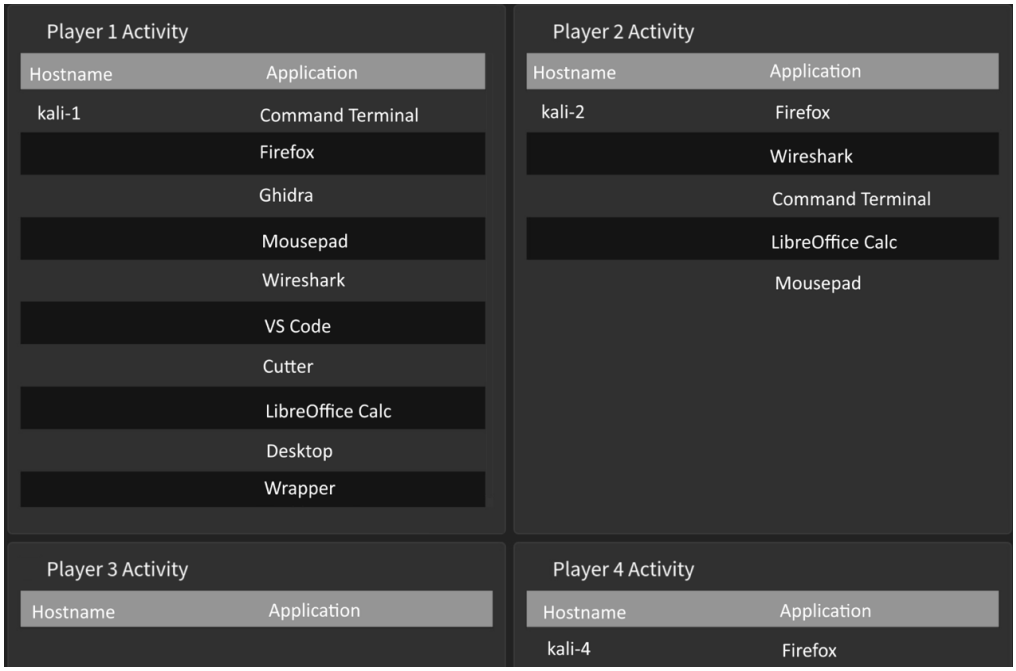

Figure 3: Comparison of Team Member Activity

5. Future Work

To date, we have collected over 10 billion lines of logs—more than 2 TB of data—from on-demand cybersecurity exercises deployed for the President's Cup Cybersecurity Competition. In Section 4, we showed a brief analysis of this data, highlighting a few aspects of how command history and graphical application usage data can be used to draw conclusions about user experience and to map tool usage to NICE work roles. There is significant future work to be done to conduct a more detailed and thorough analysis of this data.

One path for future work is to map software tools that are used by experts to the NICE work roles that use those tools. By looking at the command history and application usage of users who have completed cybersecurity exercises, we can identify sets of tools that are commonly used to complete similar tasks. We began this analysis in Section 4, but a more complete analysis would result in a taxonomy that links tools used by experts to NICE work roles.

Another path for future work is to develop predictive models and data-driven approaches for evaluating cyber talent. In a data-driven evaluation of cyber talent, we would provide users with a cybersecurity exercise, then correlate their activity data with their exercise results (i.e., successful completion, partial completion, or no completion). Based on the results of the exercise and how the user interacted with various tools, we could make an inference on the user's expertise. For example, if User 1 completes an exercise and their activity data shows that they are using tools as designed, we can infer that the user has expertise in the topic area. Similarly, if User 2 partially completes or does not complete the exercise and their activity data shows that they are trying several different paths without arriving at a solution, we can infer that User 2 might be less experienced in this topic area than User 1.

A third path for future work could integrate artificial intelligence (AI) with the data that we have along with additional data we could collect. One example of this would be to train AI models using the user activity data that we have collected. With the AI model trained on how users approach cybersecurity exercises, it might be possible for the AI system to generate a solution to the exercise. The AI model could be further enhanced by training it on screen recordings of users completing the exercise.

Finally, additional future work can build on the sabermetrics fundamentals – using data to improve team and individual performance. In Section 4.2, we examine data that shows how individual users approach exercises. Users that demonstrate mastery of a topic (e.g., a tool, NICE Work Role, etc.) are able to solve exercises quickly and users with less proficiency are either slower to complete the exercise or fail to complete the exercise. In Section 4.3, we briefly analyze how members of a team can work together to problem solve while working on different parts of an exercise. By combining these analyses, we can find what teams of cybersecurity professionals are good at, isolate topics where the team could use improvement, and even provide

suggestions on how to build/operate a more performant team. Isolating gap areas where individuals and teams could benefit from additional training and highlighting what practitioners are already proficient in will serve to increase the fidelity of the cybersecurity workforce by providing personalized and targeted recommendations for improvement – much like the introduction of sabermetrics was able to provide competitive advantages to the baseball teams doing more with the data available for analysis.

6. Conclusion

Activity data generated by users participating in ephemeral cybersecurity exercises has immense value for identifying effective cybersecurity workflows and evaluating cybersecurity talent. To use this data effectively, it must first be extracted from the ephemeral environment and preserved for detailed analysis.

The SEI has demonstrated a working system for log collection and storage of user activity data. We have also shown preliminary analysis of the data we have collected. Future work in analyzing our existing data and collecting additional data can increase the impact of what we have done thus far.

Acknowledgements

Carnegie Mellon University 2025

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM25-0258

References

ActivityWatch, n.d. *ActivityWatch.* [Online] Available at: https://activitywatch.net/ [Accessed August 2024].

Booz, J. et al., 2022. *Challenge Development Guidelines for Cybersecurity Competitions,* Pittsburgh: Software Engineering Institute, Carnegie Mellon University.

Cybersecurity and Infrastructure Security Agency (CISA), n.d. *President's Cup Cybersecurity Competition.* [Online] Available at: https://www.cisa.gov/presidents-cup-cybersecurity-competition [Accessed August 2024].

Elastic, n.d. *Filebeat.* [Online] Available at: https://www.elastic.co/beats/filebeat [Accessed August 2024].

Executive Office of the President, 2019. *Executive Order 13870: America's Cybersecurity Workforce.* [Online] Available at: https://www.federalregister.gov/documents/2019/05/09/2019-09750/americas-cybersecurity-workforce

Graylog, Inc., n.d. *Graylog Sidecar.* [Online] Available at: https://go2docs.graylog.org/current/getting_in_log_data/graylog_sidecar.html [Accessed August 2024].

Graylog, n.d. *Graylog.* [Online] Available at: https://graylog.org/ [Accessed August 2024].

json.org, n.d. *JSON.* [Online] Available at: https://www.json.org/json-en.html [Accessed August 2024].

Lewis, M., 2004. *Moneyball.* New York: W.W. Norton. Man7.org, n.d. *Script(1)—Linux manual page.* [Online] Available at: https://man7.org/linux/man-pages/man1/script.1.html [Accessed August 2024].

Man7.org, n.d. *Tail(1)—Linux manual page.* [Online] Available at: https://man7.org/linux/man-pages/man1/tail.1.html [Accessed August 2024].

Microsoft, n.d. *Visual Studio Code.* [Online] Available at: https://code.visualstudio.com/ [Accessed August 2024].

National Initiative for Cybersecurity Careers and Studies, n.d. *Digital Forensics.* [Online] Available at: https://niccs.cisa.gov/workforce-development/nice-framework/work-role/digital-forensics [Accessed August 2024].

National Initiative for Cybersecurity Careers and Studies, n.d. *Workforce Framework for Cybersecurity (NICE Framework).* [Online] Available at: https://niccs.cisa.gov/workforce-development/nice-framework [Accessed August 2024].

nmap.org, n.d. *Chapter 15 of Nmap References Guide.* [Online] Available at: https://nmap.org/book/man.html [Accessed August 2024].
PortSwigger, n.d. *BurpSuite.* [Online] Available at: https://portswigger.net/burp/communitydownload [Accessed 2024 August].
Sleuth Kit Labs, n.d. *Autopsy.* [Online] Available at: https://www.autopsy.com/ [Accessed August 2024].
sqlmap.org, n.d. *Sqlmap.* [Online] Available at: https://sqlmap.org/ [Accessed August 2024].
Wikipedia, n.d. *Sabermetrics.* [Online] Available at: https://en.wikipedia.org/wiki/Sabermetrics [Accessed August 2024].
Wireshark, n.d. *Wireshark.* [Online] Available at: https://www.wireshark.org/ [Accessed August 2024].