Comprehensive Data Analytics for Threat Detection

Jaime C. Acosta[a], Monika Akbar[b], Alex Fielder[a]
[a]*U.S. Army Research Laboratory*
White Sands Missile Range, NM, USA
[b]*University of Texas at El Paso*
El Paso, TX, USA

***Abstract*—**Network traffic and system logs are primary focal points for intrusion detection, but these data only tell a part of the story. Lacking are data associated with attacker actions on attacker systems that include tools used, strategies, and how these are associated to the network traffic in order to strengthen intrusion detection systems and to perform automated security testing and provide decision support.

In this paper, we present a comprehensive cyber data analysis workflow that consists of notional cybersecurity scenarios, practical exercises, data collection tools, and an analysis component. Workshops are used to host scenarios where users are tasked with scanning, detecting, and exploiting weak points in different emulated networks, hosts, and services, in order to evaluate security posture. We have applied the workflow to collect several feature-rich datasets, released as open source, that consist of raw and formatted network traffic, keystrokes, system calls, and screenshots, among others. We conclude by describing how our analysis tools can be used to further analyze these types of comprehensive datasets.

Keywords—cybersecurity; data mining; network security; network data

I. INTRODUCTION

In field of cyber security, data are a valuable, yet scarce commodity. The data that are available is commonly used for generating models for intrusion detection, and improving efficiency and accuracy of assessments, such as penetration tests. Cyber-related incidents are common and networks typically keep activity traces, however, these data are not shared among the security community due to its sensitivity. Most cyber datasets that are available to the public are collected during capture the flag competitions. However, these datasets have limitations. These mostly consist of only network data, the data are mixed (participants are on the same network), and in many of these events, the objectives are not necessarily representative of real-world scenarios and instead focus on the competitive, game, aspect in the assigned tasks.

In this paper, we introduce a workflow that allows researchers to capture realistic and comprehensive datasets. We provide several datasets that are available online[1].

II. RELATED WORK

A major focus of network security is the timely detection of malicious intrusions. Among others, machine learning algorithms have been successfully used for detecting network intrusions. Researchers noted that while no single machine learning algorithm can efficiently identify all types of attacks, random forest and decision tree classifiers perform relatively better than the other algorithms [1][2].

Actions performed by the attacker are an important element in building attacker persona or profile. Such profiles are generated and used for assessing network security risks [3][4].

Both in the case of intrusion detection systems and attacker profiling, collection methods are mostly based on network data and logs collected on a remote host, not on the attacker's machine.

III. COMPREHENSIVE DATA ANALYSIS WORKFLOW

Our dataflow is architected as several, separate, but cohesive components.

A. Components

Workshops are built based on real-world, public knowledge. Workshops consist of scenarios that run on emulation and virtualization technologies including the Common Open Research Emulator (CORE), and VirtualBox, in order to provide more realism with respect to topologies, systems, services, and network traffic. Workshop execution is monitored and managed using the Emulation Sandbox (EmuBox) due to its small footprint and scenario isolation features [5]. The Evaluator-Centric and Extensible Logger (ECEL) is installed on all virtual machines to collect user actions and network data. These data are then parsed and stored into a Mongo database that is read by the data analysis component. The workflow is shown in Figure 1. We used our workflow to create and collect data for the scenarios described in the following sections.

B. Scenario: Pivoting and Exploitation

In this scenario, participants are located on the Internet and must gain access to an email server that resides in an Intranet. The Intranet has a host that is running a publicly accessible and vulnerable JBoss service. Participants are given a document that was apparently found while dumpster diving. It describes several subnetworks in the Intranet, including IP addresses and subnet masks. Participants must gain access to a webserver on the Intranet.

Participants complete the following tasks. Find the IP address of the node serving the JBoss service by scanning the Intranet. Use Metasploit to identify and exploit a vulnerability in the JBoss service and to run a Meterpreter session. Configure the compromised node as a pivot by configuring

[1] https://github.com/ARL-UTEP-OC/ecel-datasets

routing and using a socks4a proxy. Access the internal mail server using the browser.

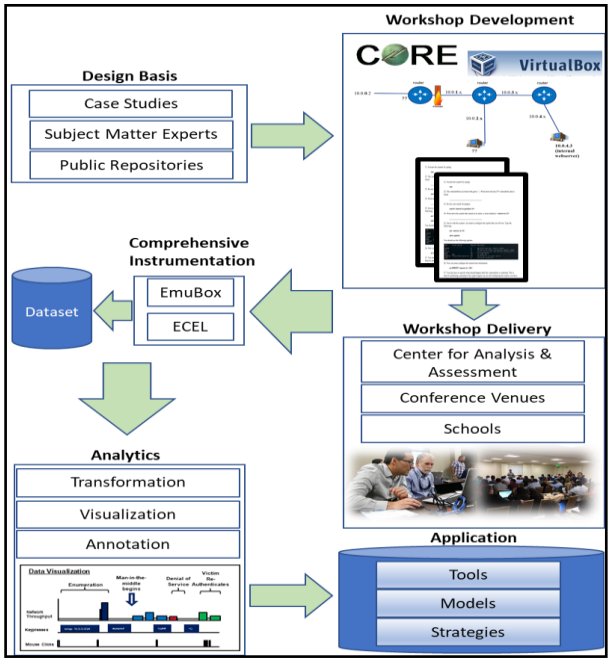

Figure 1 The comprehensive data analysis workflow

C. *Scenario B: Route Hijacking*

In this scenario, participants are not given any prior knowledge about the network. They are connected to a routing gateway that is using the Routing Information Protocol (RIP) for dynamic routing.

Participants complete the following tasks. Use Wireshark to view the routing network packets and identify all subnetworks. Spoof a plaintext authentication web page running on a remote host. Host the spoofed web page and then use the Loki.py tool to advertise a false route to the web server. Use Wireshark to retrieve user credentials.

IV. Preliminary Data Analysis

We have developed the following analysis components. The timeline viewer, shown in Figure 3, is used for editing, annotating, and extracting portions of workshop-related data. This is used to map attacker actions to network traffic and to build models for decision support and attacker profiles.

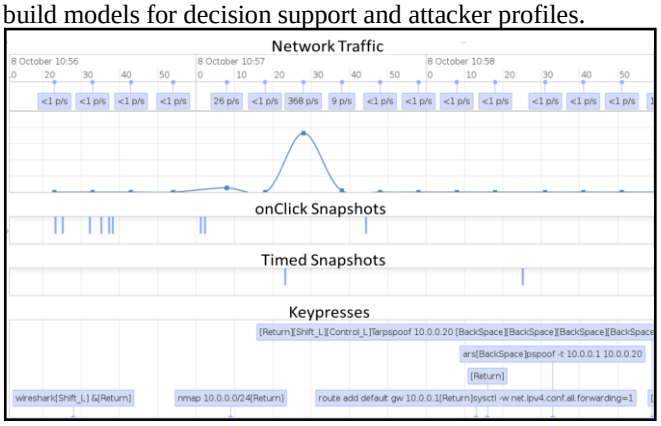

Figure 3 Timeline visualization screenshot

The heatmap viewer, shown in Figure 3, is used to identify similarity in network traffic across traffic captures and scenarios. This heatmap shows 2 captures with high occurrences of the ***tftp*** lexicon (and, hence, the protocol). This will be used to improve intrusion detection systems and also to aid during security assessments (such as penetration testing) and to fine-tune and prune attack graphs, e.g., by assigning confidence metrics based on attacker profiles [6].

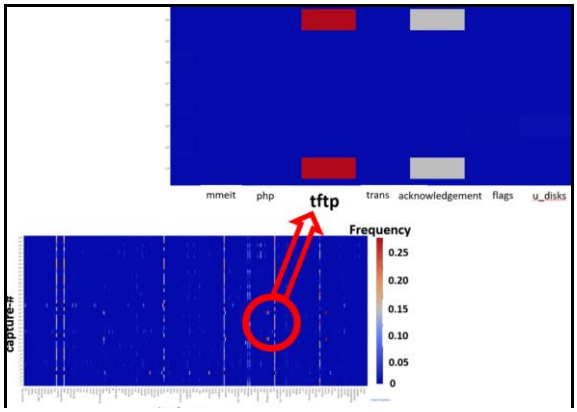

Figure 2 Heatmap for several network captures

V. Future Work

In the future, we will employ temporal pattern mining and motif mining techniques to investigate the workshop data for detecting suspicious activities that frequently appear together in attacker's data streams. We will study the correlation between network characteristics, network traffic, and system commands. We will also conduct further studies to identify the best approach for modeling attacker's profile based on pre-intrusion and post-intrusion activities at the network and system level.

References

[1] M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, 2017, pp. 000277-000282. doi: 10.1109/SISY.2017.8080566

[2] M. A. Jabbar and S. Samreen, "Intelligent network intrusion detection using alternating decision trees," *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, Bangalore, 2016, pp. 1-6. doi: 10.1109/CIMCA.2016.8053265

[3] J. Brynielsson, U. Franke, M. Adnan Tariq and S. Varga, "Using cyber defense exercises to obtain additional data for attacker profiling," *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, Tucson, AZ, 2016, pp. 37-42. doi: 10.1109/ISI.2016.7745440

[4] D, Ram, K, Prakash, C, Joao, "Network risk management using attacker profiling", 2009, Security and Communication Networks. 2. 83-96. 10.1002/sec.58.

[5] Acosta, J. C., McKee, J., Fielder, A., Salamah, S. (2017 October). A Platform for Evaluator-Centric Cybersecurity Training and Data Acquisition. In Military Communications Conferenceu MILCOM 2017. IEEE.

[6] Acosta, J. C., Padilla, E., & Homer, J. (2016, November). Augmenting attack graphs to represent data link and network layer vulnerabilities. In Military Communications Conference, MILCOM 2016 (pp. 1010-1015). IEEE.