Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets

Benjamin Sangster, T. J. O'Connor, Thomas Cook, Robert Fanelli,
Erik Dean, William J. Adams, Chris Morrell, Gregory Conti
United States Military Academy
West Point, New York

Abstract

Unlabeled network traffic data is readily available to the security research community, but there is a severe shortage of labeled datasets that allow validation of experimental results. The labeled DARPA datasets of 1998 and 1999, while innovative at the time, are of only marginal utility in today's threat environment. In this paper we demonstrate that network warfare competitions can be instrumented to generate modern labeled datasets. Our contributions include design parameters for competitions as well as results and analysis from a test implementation of our techniques. Our results indicate that network warfare competitions can be used to generate scientifically valuable labeled datasets and such games can thus be used as engines to produce future datasets on a routine basis.

Keywords: DARPA dataset, Lincoln Labs dataset, Cyber Defense Exercise, Capture the Flag

Background and Motivation

Capturing network traffic is a relatively straightforward process, but raw network traffic data is of limited value to researchers seeking to test network security techniques, notably intrusion detection systems. A more useful network traffic capture is one in which the dataset traffic is labeled in some suitable fashion to support security analysis. To address this shortcoming, the Defense Advanced Research Projects Agency (DARPA) and the Massachusetts Institute of Technology's Lincoln Labs partnered to produce the DARPA 1998 and 1999 datasets. [1,2] The datasets contained labeled data generated by simulating network traffic for a medium size U.S. Air Force base. While the datasets included some known shortcomings such as relatively low traffic rates, lack of realistic Internet background noise, and lack of validation [3], the datasets were highly innovative for their time and were widely used by security researchers. However, the utility of the DARPA datasets has declined over time due to aging content and continually emerging threats, eventually reaching a point where many researchers avoid publishing results solely based on the datasets. [4] No modern replacement exists for the DARPA datasets.

To help fill this gap, we propose techniques for strategically instrumenting network warfare competitions to generate scientifically valuable labeled datasets. Dozens of network warfare competitions occur on an annual basis and many contain key elements required for useful dataset generation, including defenders (blue) and attackers (red) as well as traffic generation (white). By strategically placing network sensors as close as possible to data generation it is possible to automatically characterize much of the network traffic as red, white, or blue, a significant advance from naive approaches that capture only a mixture of traffic at a single, centralized collection point. In addition, by gathering additional semantic information from the competition, such as configuration data, hard disk images, and device logs, it is possible to augment traffic captures with a rich set of supplemental information. To date, network warfare competitions have not been explored as mechanisms for generating labeled datasets.

While our results indicate this approach is viable, we do not claim current instrumentation techniques can surpass the value of the 1998 and 1999 DARPA datasets at the time of their release. However, we argue the techniques we propose can provide a higher quality of dataset than current naive data collection provides and that with future work significantly higher value dataset generation is possible on a routine basis.

We make several contributions in this paper. First, we compare popular game architectures to explore the impact of game design on projected dataset output. We then provide results and analysis from a network instrumentation and data capture experiment conducted during the 4-day Cyber Defense Exercise competition between the U.S. Military Academy and a National Security Agency (NSA) red team.

This paper is organized as follows. Section 2 studies popular game architectures and proposes suitable formats for generating datasets. Section 3 discusses the structure, execution and results of an experiment conducted to test our approach. Section 4 analyzes our results and Section 5 presents our conclusions and promising directions for future work.

Competition Characteristics

Network warfare competitions come in many varieties, each with differing likelihood of generating useful datasets. The most distinguishing characteristic of network warfare competitions is their offensive or defensive nature. Offensive competitions are games in which teams attack each other and seek to fend off incoming attacks, or attack a common set of targets which have associated point values. In offensive competitions there is typically no formal red team, instead each individual team includes integrated red (offensive) and blue (defensive) roles. In addition, teams may be required to provide network services such as chat, email, and web serving which are monitored for up-time by judges. The canonical examples of this class of competition are Defcon's Capture The Flag (CTF) competitions run by the Kenshoto (kenshoto.com) and Ghetto Hackers (www.ghettohackers.net) groups. Defensive competitions differ in that teams are prohibited from attacking other teams and instead must secure networks and often must provide consistent network services. Offensive activities are allowed only by officially sanctioned neutral red team members. Examples of defensive competitions include the NSA-sponsored Cyber Defense Exercise (CDX) and the National Collegiate Cyber Defense Competition (NCCDC). [5, 6]

The type and nature of the traffic generated during the competition, whether red (offensive), white (simulated end user traffic), or blue (defensive) is critically important to the quality of the resulting dataset. The ideal goal from the data capture perspective is to generate network traffic that accurately emulates real world traffic, but is generated by hosts that emit primarily red, blue, or white traffic. This traffic can then be coarsely identified as primarily red, blue or white by collecting the traffic at nearby network sensors. However, many factors influence this outcome. Both humans and machines may generate red, white, or blue traffic. When humans generate traffic, their skill level and the tools they employ dictate the quality of the output. In the case of red traffic, novice participants may generate traffic that contains already solved detection problems, such as network or vulnerability scans or common automated attacks like those generated by Metasploit or Core Impact. Advanced users however, may attack in more subtle ways and could utilize exploits that have not been disclosed publicly. Games may also include white traffic generators. Commonly these generators are automated, employing anything from simple scripts to perform tasks like downloading web pages to sophisticated traffic generation tools such as LARIAT which can be configured to simulate large numbers of users performing a variety of activities. [7] Sophisticated traffic generation tools are powerful, but ultimately depend on their underlying models for realism; the more accurate the model, the more realistic the traffic generation. A significant advantage of automated traffic generation is that the traffic generators can automatically create precise logs that may be directly correlated to their network traffic.

While automated white traffic generation is relatively common, some competitions, such as those conducted by White Wolf Security (www.whitewolfsecurity.com), include traffic generated by human white teams. Human generated white traffic is intriguing because of its lack of dependence on computer models and its potential for realistically generating legitimate traffic. However, incorporating human generated traffic is resource intensive, requiring many hours of human effort, and is difficult to scale to levels of automated traffic generation systems. This point leads to the issue of traffic volume and proportion.

Existing network warfare competitions are typically short duration events, on the order of one to four days. They tend to contain roughly an equal ratio of malicious and non-malicious traffic. Contrast this balance with real-world network conditions where one would expect a far higher proportion of legitimate to malicious traffic. Real-world malicious traffic would contain automated attacks (including modern worms such as Conficker and legacy worm activity such as Code Red) and, rarely, human attackers. Due to the toxic nature of network warfare competitions, many are conducted on air-gapped, or otherwise isolated networks, and are thus not conducted directly on the Internet. Because of this, Internet background radiation caused by automated worms, malformed packets, flooding backscatter, etc. [8] will largely be absent in network warfare competition datasets. The end result is that current competition designs will have a disproportionately high volume of malicious activity and a disproportionately low volume of non-malicious activity.

The support of participants is also significant. Some competitors may not wish to have their online activities monitored and captured. Others may support network collection, but refuse to allow any form of collection that requires instrumenting their individual workstations. In addition, participants may resist collection if they perceive the information would give any team an unfair advantage.

Individual roles within teams are also important, particularly in cases where teams mix offensive and defensive roles. Some teams might assign individual members a purely offensive or defensive role, allowing for potentially straightforward collection and labeling of traffic. Other teams, however, might organize such that individual players rapidly alternate between

offensive and defensive roles, resulting in mixed red and blue traffic which frustrates automated labeling.

Each competition follows a specific set of rules established by organizers. Rule sets vary, but they typically dictate hardware, software, operating systems, and services that must be provided (or are prohibited) as well as stipulate the composition and general mission of participating teams. In addition, rules can influence game play in ways that may constrain network topography and inhibit the ability of placing network sensors. Hence rules fundamentally impact the type and quality of data that will be created in the course of a game. Today's competitions are not designed to generate labeled datasets. However, future competitions may be designed to support dataset generation. We will discuss this issue later in the paper.

Collection Architecture and Instrumentation

To test the efficacy of our approach, we carefully instrumented the 2009 Inter-Service Academy Cyber Defense Exercise. The 2009 CDX was a complex 4-day exercise which incorporated a professional NSA red team and automated white traffic generation.

For this test we built three collection systems (listed below) from commodity personal computer components each running FreeBSD 7.1. We used TCPDUMP for packet capture.

A. Dual 2.33GHz Xeon Quad-Core Processors, 24GB RAM, 2.5TB RAID-5 Secondary Storage (FreeBSD 7.1 amd64)
B. Dual 2.33GHz Xeon Quad-Core Processors, 24GB RAM, 5TB RAID-5 Secondary Storage (FreeBSD 7.1 amd64)
C. Four 2.7GHz Xeon Processors, 24GB RAM, 365GB RAID-5 Secondary Storage (FreeBSD 7.1 i386)

The CDX was held 21-24 April 2009 and consisted of teams from the U.S. Military Academy (USMA) and seven other military colleges. These teams built and secured operational networks within constraints specified by the exercise. Each team built their network from trusted operating system distributions (both Linux and Windows-based), but was required to integrate three untrusted workstations which were provided by the exercise organizers as VMWare images. As part of the competition, each team was tasked to provide consistent network services including a web application that included database-driven dynamic content, chat services utilizing the XMPP protocol, email, Domain Name System (DNS), and Microsoft Active Directory. The networks were then attacked by an NSA red team consisting of approximately 30 personnel. To mask their attacks the red team generated non-malicious Simple Mail Transfer Protocol (SMTP) and HyperText Transfer Protocol (HTTP) traffic using a simple traffic generation program. Participating college teams were forbidden from conducting offensive operations. Networks were subject to attack from 0800-1600 on each of the four days of the competition. The exercise was conducted over a virtual private network (VPN) connecting each participating team. No external Internet traffic was allowed into the VPN and thus Internet background radiation does not appear in the resultant dataset.

In addition to the red and blue teams traditionally associated with cyber warfare exercises, other members of the NSA served as a neutral judging cell. At each team location, a judge was present to validate service functionality throughout the exercise, as well as to supervise compliance with the CDX rules. Electronic communication between judges is present in the final dataset.

We placed network sensors at three locations during the exercise as shown on Figure 1. The first sensor (A) was placed on the network path connecting the Red team to the exercise via a SPAN port on a Cisco 2811 router. All red team initiated traffic was visible to this sensor. The second sensor (B) was placed on the network connection just inside the USMA team's VPN router via an inline passive tap. The final sensor (C) was placed at a central location inside the USMA network perimeter at the primary network switch, a Cisco 2960G. Based on their locations, sensors B and C monitored a mix of red, white, and blue traffic, but Sensor A monitored primarily red traffic. Connection of Sensor B was achieved using a NetOptics (www.netoptics.com) passive monitoring tap (10/100/1000BaseT). Sensors B and C were placed to observe the traffic on each side of the USMA network perimeter. Sensor B captured all inbound traffic while Sensor C captured the inbound traffic remaining after ingress filtering. Similarly, Sensor C captured all outbound traffic while Sensor B captured the traffic remaining after egress filtering. Further, the traffic captured at Sensor C was identical to that visible to the USMA network intrusion detection system during the exercise. To scope the experiment, we chose to instrument only the NSA Red team and the USMA network as part of this experiment and leave full instrumentation of the other teams participating in the CDX as future work. However, we believe our placement of sensors was sufficient to validate our instrumentation approach.

It is important to note that NSA provided explicit permission for us to collect data under the condition that we did not use the data to influence the exercise in any way. We strictly abided by this guidance. Moreover, NSA approved of sharing the data to the public after completion of the CDX.

In addition to numerous manual attacks during the course of the exercise, we observed a denial of service attack launched by the NSA red team consisting of a 15 minute high-volume SYN-flood attack. We also observed near constant automated scanning of hosts and

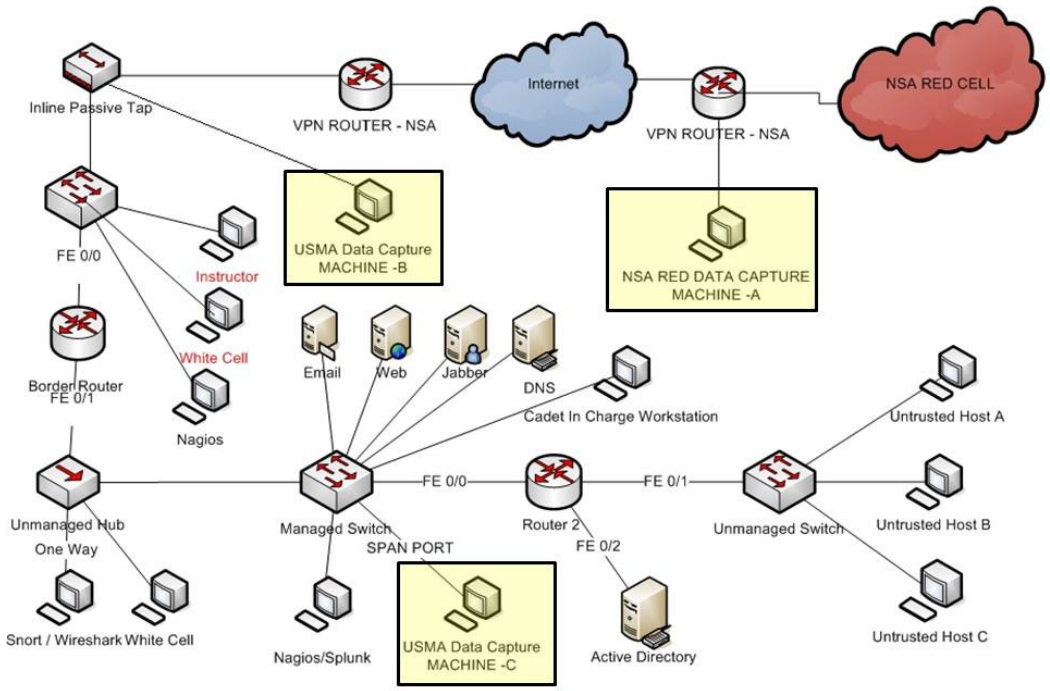

Figure 1. Instrumented portion of the 2009 Inter-Service Academy Cyber Defense Exercise network with capture nodes highlighted.

services in an attempt to enumerate the USMA network. Despite repeated attacks by the NSA red team targeting email, web, and DNS, and other services, the USMA team did not suffer a major compromise.

Analysis

The labeled DARPA datasets of 1998 and 1999 were the security community benchmark for testing intrusion detection systems. After ten years, these datasets contain significant limitations. In this section, we examine the role network warfare games can play in augmenting these de facto evaluation datasets by studying the results of our CDX instrumentation experiment.

Traffic Artificiality

The artificiality of the labeled DARPA dataset is a shortcoming that was raised by McHugh. He argued that the traffic generation methods used to make the datasets created unintentional artifacts that differentiated malicious and benign traffic [3]. To demonstrate this, Mahoney and Chan constructed a trivial intrusion detection system by noting that all malicious packets in the DARPA dataset had a TTL of 126 or 253. [9] Although a valuable effort, the DARPA data sets lack some of the attributes seen in traffic generated interactively by live human users and attackers. In situations where human traffic is not possible, the LARIAT test bed for traffic generation has shown real promise by generating simple and multi-phased attacks against defensive technologies, while sustaining high traffic rates.

The dataset recorded from the CDX contains a significantly different personality than the scripted DARPA dataset. Although attackers used tools such as Nessus, WebScarab and Nikto to automate reconnaissance and attacks, the overall generation of attack traffic was human directed and implemented by a 30 person team from the NSA. Using actual network warfare game players reduces the artificiality of the CDX dataset. Concurrently, a 20 person team of game organizers generated white traffic by manually interacting with web, email, DNS lookups, and other required services. Internal to each team's network, three virtualized workstations, generated traffic from a series of Ruby scripts that crawled across web pages on the domain. The mixture of cover traffic with malicious traffic raises an important issue for IDS researchers because the mixture inhibits clearly labeled red traffic. We believe this shortcoming can be overcome by gathering detailed red team log data which could be correlated with packet captures. We will cover this point in more detail in the next section. In addition, IDS researchers may find it possible to train systems by treating the combination of cover traffic and malicious traffic as a unified and detectable event.

A weakness of current network warfare games such as the CDX is the lack of the volume and diversity of traffic normally seen in production networks. Due to the potential risks both to the Internet at large and to the orderly conduct of the exercise, these games are normally conducted on isolated networks. Therefore, these exercises lack typical Internet background noise. This presents an issue as Paxson argued that large volumes of legitimate and benign network traffic exhibits abnormal behavior such as FIN/RST floods, private IPs leaking onto the public Internet, or fragmented packets with the Don't Fragment bit set [10]. In order to achieve a more realistic dataset, it is ideal if future network warfare games be played on the actual Internet. We believe this goal may be achievable, some online competitions have taken place, but future large scale competitions, particularly those instrumented to collect full packet payloads, require significant research into risk mitigation and privacy protection. However Honeypots may provide a viable compromise. For the past decade, The HoneyNet Project has used many Honeypots to successfully generate mostly red datasets that have proven successful in helping analysts learn more about such attacks as the Conficker worm. The recent propagation of worms such as Conficker have demonstrated the necessity to include automated worms into future network warfare games. The dataset released from the CDX contains traffic from machines infected by custom rootkits built by the NSA specifically for the game, but lacks any actual infestation of a real-world malicious worm such as Conficker or Code Red. Future network warfare games could enable live infestations of network worms or mix in traffic captures from the HoneyNet Project or other sources.

Scale

The DARPA data sets represent the traffic of a relatively small network of 33 live and simulated hosts interacting with a total of 12 external hosts. The CDX dataset increases the scale of the network by demonstrating attack attempts from a 30 person red team using IP addresses from a pool of over sixty-five thousand host addresses against workstations, network devices, internal web servers, domain name servers, email servers, and chat servers from the 9 different collegiate team networks. The types of attacks employed against each server are significantly varied as a DNS cache spoofing attempt demonstrates different anomalous behavior when compared to a web-forgery attack. As the distribution of services between multiple machines has increased over recent years, it is necessary that network warfare games accurately represent this scenario. However, the time scale of the CDX dataset is limited to a four day exercise through a VPN. This is consistent with the typical network warfare game, lasting a week or less, during which the intensity of activity by defenders and attackers alike is significantly higher than in most production environments. The limited duration of competitions and periods of attacker inactivity are issues of concern when testing anomaly detection systems that require a training period.

We believe that future network warfare games can include extended network and time scales. For example, the NCCDC contains a series of eight different regional competitions, each lasting a week and including emerging technologies such as ecommerce servers and workstations with multiple vendor operating systems. It is entirely possible that a similar game could be played across the Internet for months with regional, national, and international competitions.

Supporting Artifacts

To augment the raw IPv4 data captures, the DARPA dataset provides Solaris Basic Security Mode audit data for the SolarisOS, file-system dumps from each day of the exercise, and process output generated every minute of the exercise. This additional data allows intrusion detection researchers to understand how network traffic affects behavior on the targeted machines. Network warfare games represent the dynamic methods in which attackers attempt to compromise a network. For example, an attacker may compromise a low value machine on the network that has a trust-relationship with a higher value machine such as the domain controller. This type of attack is clearly a call for recording data from multiple sensors as well as augmenting those sensors with host-based and network-device logs. Identifying multiple vantage points for recording traffic is essential to creating a useful dataset. While the CDX dataset has a limited set of vantage points, recording future network warfare games allows the possibility of multiple recording sensors at critical network locations. In particular, we recommend placing sensors as close as possible to sources of red, white, and blue traffic generation as well as placing sensors at central locations on the network to capture mixed traffic. Aggregating logs from all machines on a network would further assist researchers in analyzing specific attacks.

The CDX dataset provides logs aggregated from network monitoring devices, hosts, and servers on the internal USMA competition network. In addition to the actual logs, we provide detection logs of malicious traffic recognized by our internal intrusion detection system and the email traffic from the USMA Team Captain to CDX judges containing information on suspected attacks and their originating IP address. Future network warfare games could provide full-spectrum data such as detailed data on the nature, source, destination, and timing of each attack event. In

addition, virtual machine images could be captured of all devices and shared with researchers, barring copyright and privacy issues. By creating multiple snapshots of the virtual machines, intrusion detection researchers could develop a better understanding of how and when the attacks specifically compromised a device. Ultimately, future network warfare games provide real promise for recording datasets to augment the current de facto standards.

Conclusions and Future Work

Network warfare games can produce automatically labeled network security datasets, but the quality of the result depends on the structure and conduct of the game, network topology, and sensor placement. However, over time, we believe it is possible to move toward games that provide increasingly realistic network traffic by altering competition rules and providing appropriate incentives (and disincentives) to encourage players to behave as they would when conducting live attacks on the Internet. Realistic game scenarios combined with strategically placed network sensors can produce coarsely grained labeled data. This data is valuable, but more valuable fine grain labeling requires semantic information not available from network sensors alone. An interim solution, and one that we chose to implement, was to capture as much of this semantic information as possible and make the results available alongside the dataset. For future work we recommend perusing automated solutions that, when used in conjunction with strategically placed sensors, will automatically gather as much semantic information as possible. In particular, we recommend exploring ways to instrument operating system distributions, particularly those used by red teams, to generate detailed logs of exact tools used and commands issued. Detailed red team logs, whether manually or automatically generated, would provide information critical to more precise labeling of red traffic than what is possible with network sensor placement alone. In addition, we recommend pursuing machine learning techniques, perhaps trained with human assistance, that can assist in correlating host-based data with packet captures. Despite these current shortcomings, we believe that network warfare competitions are a viable solution today for creating useful datasets and supporting semantic data and bear even greater promise for the future.

.

Acknowledgements

We would like to thank the following for their support, helpful ideas, and feedback: Army Research Labs, Michael Collins, Robert Cunningham, Carrie Gates, FLOCON, Richard Lippmann, Lisa Marvel, MIT Lincoln Labs, John McHugh, NSA, and Tamara Yu.

References

1. D. Fried, I. Graf, J. Haines, K. Kendall, D. Mcclung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, M. Zissman. "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation." DARPA Information Survivability Conference and Exposition, 2000.

2. J. Haines, D. Fried and J. Korba. "Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation" Recent Advances in Intrusion Detection (RAID), 2000.

3. J. McHugh. "Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory." ACM Transactions on Information and System Security, Vol. 3, No. 4, November 2000, pp. 262-294.

4. T. Brugger. "KDD Cup '99 dataset considered harmful." White Paper, University of California Davis, Department of Computer Science, 15 September 2007.

5. "West Point Takes the NSA Cyber Defense Trophy for the Third Straight Year." National Security Agency Press Release, 28 April 2009.

6. National Collegiate Cyber Defense Competition Official Website. http://www.nationalccdc.org/

7. T. Yu, B. Fuller, J. Bannick, L. Rossey, and R. Cunningham. "Integrated Environment Management for Information Operations Testbeds." Workshop on Visualization for Computer Security, 2007.

8. R. Pang, V. Yegneswaran, P. Barford, V. Paxson, L. Peterson. "Characteristics of Internet Background Radiation." ACM Special Interest Group on Data Communications Conference (SIGCOMM), 2004.

9. M. Mahoney and P. Chan. "An analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection." Recent Advances in Intrusion Detection, 2003.

10. V. Paxon. "Bro: A System for Detecting Network Intruders in Real-Time." USENIX Security Symposium, 1998.