# Labeling the National Collegiate Cyber Defense Competition Dataset for Cybersecurity Research

Chelsea Hicks
Department of Information Systems & Cyber Security
University of Texas at San Antonio
San Antonio, Texas, U.S.A.
Chelsea.Hicks@utsa.edu

Nicole L. Beebe, Ph.D.
Department of Information Systems & Cyber Security
University of Texas at San Antonio
San Antonio, Texas, U.S.A.
Chelsea.Hicks@utsa.edu

***Abstract*—In this paper, we discuss a semi-synthetic dataset hosted by the Department of Homeland Security (DHS) Cyber Security Division IMPACT website. We contend that this understudied and underutilized dataset can overcome some of the significant challenges posed by the nearly two decades old datasets that continue to be widely used in intrusion detection and prevention research today (DARPA 1998, DARPA 1999, KDD 199 Cup intrusion detection datasets). With the DHS dataset in mind, we introduce a methodology to establish ground truth in noisy, full packet capture datasets involving multiple computers, networks, threat actors, and attacks. Last we discuss our proposed methodology to autonomously detect and characterize risk in on-going attacks, trained using this dataset.**

***Keywords—cyber security, big data, intrusion prevention system, DHS IMPACT, NCCDC dataset***

## I. Introduction

Big data and analytics are very popular research topics today, across a wide range of domains. Advances in computing storage and power now enable us to parse through datasets that were previously too large to analyze with advanced analytics techniques. As a result, there has been a call for more research in various domains, including cybersecurity, that utilizes big data and machine learning. Doing so leads to greater intelligence and better systems. However, researchers continue to suffer from a dearth of large, realistic, open-source datasets for such research. While organizations can examine their own network traffic and internal data, academic researchers often lack access to such data due to privacy and business sensitivity concerns. Further, even when researchers are fortunate enough to gain access to such government and industry datasets, the research is not reproducable, nor readily extensible by other researchers. As a result, researchers continue to rely on publicly available datasets that are often grossly outdated and/or are not representative of modern network intrusions.

In this paper, we outline a dataset that contains more sophisticated attacks than prevailing, publically available datasets. It is large, contains both benign and malicious traffic, and is only partially synthetic. However, the dataset is unlabeled and comprehesive ground truth analysis has yet to be performed and made publicly available. As such, we outline a proposed methodology to establish ground truth. Our goal is to release our ground truth analysis and labeled dataset, enabling others to utilize it.

The rest of this paper is outlined as follows. First, present a literature review of popular datasets used in cybersecurity research, previous efforts to help speed up ground truth assessment, and previous autonomous intrusion detection systems that have been developed using such datasets. We then outline the dataset and provide an overview of the scenario used to create the dataset, to provide readers an insight into what to expect from the dataset. We then introduce our methodology to establish ground truth, and we conclude with future research goals.

## II. Related Works

### A. Previous Datasets

Intrusion detection/prevention research has relied heavily on a few datasets that have been extensively used and cited, since their inception in the 1990's: 1) the 1998 DARPA intrusion detection dataset [1], 2) the 1999 DARPA intrusion detection dataset [2], and 3) the 1999 KDD Cup dataset [3].

In the DARPA 1998 intrusion detection dataset, a testbed was developed that generated normal traffic similar to a government site containing hundreds of users, on thousands of hosts [1]. In this dataset, only six research groups participated in a blind evaluation and results were analyzed for probe, denial-of-service (DOS), remote-to-local (R2L), and user to root (U2R) attacks [1]. The 1999 DARPA intrusion detection dataset consists of eight sites participating in an off-line intrusion detection evaluation [2]. In this evaluation, there was a testbed that generated live background traffic and simulated a government site containing hundreds of users on thousands of hosts. The dataset consisted of more than 2,000 instances of 58 attack types, launched against UNIX and Windows NT hosts [2]. This dataset consisted of old probing and denial-of-service attacks, user-to-root attacks, remote-to-local, and data attacks [2]. In both datasets, each research group tested their intrusion detection systems to find what attacks they could find.

In the 1998 DARPA dataset, detection rates ranged from 63% to 93%, with a false alarm rate of 10 false alarms per day. Detection rates were worse for new and novel R2L and DoS attacks. The best systems failed to detect roughly half of the new attacks, which included access to root-level privileges by remote users, which indciated a need for further research in this type of attack [1]. In the 1999 DARPA dataset, false alarm rates remained low (less than 10 per day), but range of detection rates varied greatly. One of the major aims of this dataset was to create a dataset for testing intrusion detection models built using the 1998 DARPA datset. The dataset was also created to determine these models' ability to detect new attacks that did not exist in the trainig set. Another benefit of this dataset over its predecessor was the fact it contained more sophisticated attacks. The detection rate on this dataset ranged between 53% and 100%, depending on the type of attack and what tool was used [2]. Therefore, when comparing the two datasets, the 1999 DARPA dataset and evaluation shows an overall improvement in detection rates among these attacks. It is important to note that some of the newer and more sophisticated attacks in this dataset were entirely missed by all detection systems. However, there was obvious improvement in the intrusion detection systems, showing that research was progressing the correct way.

A third, widely used dataset is the KDD 1999 Cup Dataset. This dataset was used for the Third International Knoweldge Discovery and Data Mining Tools Competition. The task was to build a network instrustion detection system that used a predictive model to distinguish between instrusions/attacks and good connections [3]. The dataset contains a standard set of data to be audited, which includes a variety of instrusions simulated in a military network enviornment (inspired by the 1998 DARPA intrusion detection dataset). Like the previous datasets, the KDD 1999 Cup dataset attacks fell into four categories: denial-of-service, R2L, USR, and probing attacks. Thus, while this dataset was new, it did not provide materially new types of attacks, or anything else significantly different from the other two datasets.

These three datasets remain among the most widely cited for intrusion detection research still today, which we find disconcerting. Other than the sheer old age of these datasets, the attacks they contain are trivial relative to today's advanced attack techniques. While some of these attacks still occur, such as the case when organizations continue to use legacy hardware, they are no longer representative of the threats organizations must be worried about today. For example, these datasets do not contain phishing or other social engineering attacks, nor advanced persistent threats, nor insider threats, nor the use of modern equipment. Overall, the sophistication and the nature of the attacks themselves have changed.

The cybersecurity field is in dire need of new, rich, publicly available intrusion detection datasets. The field needs more representative datasets to ensure research and systems developed using them are relevent and valid [4]. It is both difficult and time consuming to obtain such datasets [4]. Further, such datasets are largely unlabeled, making the use of machine learning and data mining methods challenging. The dataset introduced in the next section, coupled with the proposed ground truth analysis and labeling procedure will help close the current intrusion/prevention dataset gap.

### B. *Previous Efforts to Establish Ground Truth*

One of the distinct benefits of the previously criticized datasets is their labeled nature. Labeled datasets, often established through post-hoc, manual ground truth analysis, are critical to machine learning based approaches—both in the training of supervised learning models, as well as in the validation of unsupervised learning models. Ground truth is the act of going through a dataset to determine what happened. In the case of labeling, this may consist of identifying network traffic as malicious or not malicious, and what type of malicious attack was it. For example, in the previous datasets discussed in this paper, ground truth was established and four high level malicious attacks were labeled: DOS, R2L, U2R, and probes. More granular labels were then assigned thereafter.

Ground truth analysis remains a very manual, laborious, and error prone task task [5]. Being manually derived, ground truth labels are subjective and prone to vary between analysts. As such, there have been research attempts to automate the process of establishing ground truth. Therefore, there is a huge desire for an automated, or at least semi-automated, way to establish ground truth in a dataset. From this stream of research, one particular system stands out: the Ground Truth Verification System (GTVS) [5].

It is important to first note that GTVS was not created to replace the manual verification process, but instead was designed to accelerate it [5]. It was designed with two principles in mind: efficiency and accuracy. The principle of efficiency is to always try to work on higher aggregations of network flows whenever possible, such as a service level rather than individual flow. The principle of accuracy is to only make decisions with very high confidence [5]. In their report, the authors show how to use GTVS in a 30 minute trace of how to speed up the process of ground truth. Their methodology at a high level is as follows: using signatures, if they find matching results for a specific end point that appears to be strongly consistent, they assume that they have identified a particular service on that endpoint. Initially, thresholds should be set in a conservative way (especially for subsets of signatures missing, such as HTTP versus BitTorrent), and relaxed in future iterations [5]. The next step is to derive information from hostnames, such as identifying HTTPS traffic to particular services, and so on. The next step is to lower the thresholds of signatures to see if additional insight could be gained, and doing some manual inspection to verify conclusions drawn. Then behavioral characteristics of hosts with regard to overlay networks are examined, to help differinate behavior such as SMTP traffic versus peer-to-peer behavior. Finally, the last step is to further exampine the traffic manually, which they argue is feasible given flow sizes remaining [5].

While the GTVS methodology is promising and useful, the signatures that the authors reference are vague. This is due to the fact that they developed a system to help with this that contains these signatures. However, in our experience, this system is no longer maintained and cannot be set up – which makes using these signatures impossible.

### C. *Previous Autonmomous Intrusion Detection Systems*

As stated earlier, attacks have changed in style and sophistication since the popular DARPA and KDD datasets of the 1990's. One of the largest threats to organizations these days is the advanced persistent threat (APT). This is a sophisticated attacker, who is not randomly attacking wherever they see an easy opportunity, but is instead targeting an organization for a specific reason and who is willing to go through extra measures to maintain persistence on and access to compromised systems and networks. Thus, they are more dedicated, and usually more skilled. APTs are typically either supported by nation states, or are engaged in organized cybercrime, financial crimes, espionage, and terrorism [6]. Traditionally, APTs follow a cyber kill chain that is as follows: 1) reconnaissance, 2) weaponization of an infection vector, 3) exploit delivery, 4) vulnerability exploitation, 5) persistence establishment, 6) command and control set-up, and 7) attack goal attainment [6] [7]. APTs have advanced skill to thwart detection and achieve their objectives. This makes finding them difficult, and the damage potential they have is enormous. Hence, finding APTs is a hot research topic, one we are interested in solving in an autonomous manner, and one that requires more advanced and labeled intrusion detection datasets.

Recent attempts to autonomously detect APTs leverage correlation frameworks that consider the various stages of the cyber kill chain, including intelligence gathering, initial compromise, command and control, lateral movement, asset and data discovery, and data exfiltration [8]. These approaches rely heavily on: 1) blacklists of known malicious domains, IP addresses, and SSL certificates; 2) detecting use of anonymous networks, and 3) inspecting DNS traffic [9]. In doing so, these approaches fail to be truly autonomous and rely heavily on continuous rule set updating, signature analysis, and manual analysis. Other approaches introduce state look-ahead trees to recursively compute future attack decisions based on the current attack step options, providing unique viewpoints [9]. However, the graphs must be created by a security analyst manually, which limits the scalability of such approaches.

The creation of a more modern, labeled intrusion detection dataset is motivated by our desire to leverage machine learning of actual attack flows to learn how to differentiate attackers of differing sophistication levels. Others have integrated intrusion detection with active response, but they have struggled to properly attribute network events to a specific attack in networks and systems with multiple compromises (either by multiple sources for differing objectives, or by singular sources for the purpose of attack intent and/or source obfuscation) [10]. Still others have incorporated security measures into the attack graph calculus, thereby better assessing risk by incorporating vulnerability and likelihood of occurrence based on architectural defensive strength [11], [12]. Our work builds upon these related works and our past work [13], by incorporating cyber attack lifecycle patterns and producing a quantified risk assessment, with the long-term goal of supporting autonomous, active, defensive system response.

## III. Proposed Dataset

Now that we have established the need for a new, current, labeled intrusion detection/prevention dataset, we introduce the proposed dataset. We did not develop the proposed dataset. It is the full packet capture of attack, defend, and operational network traffic for the National Collegigate Cyber Defense Competition (NCCDC). The dataset consists of PCAP data for six different years: 2008, 2011-2015. The dataset is stored and available for download (after registration and request validation) through DHS's Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) [14] (previously known as PREDICT).

NCCDC is a collegtiate level competition, as the name implies. It is a competition consisting of three different stages generally: an online virtual qualification, a regionals qualifiers, and finally the national level. These different qualificaiton rounds exist due to the amount of interest that this competition has, which is normally more than 200 colleges. We highlight this fact as we believe this competition dataset can be used to help show different attack sophitication levels, such as the APT threat, whereas the DARPA and KDD datasets cannot. Since the best teams of the nation compete, and some of the best penetration testers attempt to stage different types of attacks against the teams along side more junior red-teamers being apprenticed. Several of the advanced attack techniques employed by the more senior red-teamers are in line with modern APT threats.

Each team, consisting of eight college students, participates in a two-day scenario in which they are the new information technology team hired for an organization. They must engage in system and network discovery and mapping, vulnerability assessment and risk mitigation, and defend the systems and network from a wide range active attacks—all while maintaining business continuity, operations, and responding to user and organization requests. Each of these student teams comprise the "blue team." The active attacks occur from the "red team," which consist of some of the industry best penetration testers and hackers, who dedicate their time and effort to provide a fair and equal experience to each team in their region.

In the NCCDC dataset provided through IMPACT, all that is provided is a very brief overview of how the data was collected. It is mentioned that a star network topology is used, and where each team and major group (such as the red team) are connected to a core switch. The logs were captured from this core switch thorugh the span port. As a result, at least for 2015, over 1 TB of data was captured over two days. They then provide an overview of the infrastructure each blue team had to protect and defend. There are only ten finalist teams represented in th NCCDC dataset, each defending an identical network. The red team verifies that each team starts off with the same attacks, and based on the defensive posture, the attacks mutate to each specific team.

We contend that while this dataset is semi-synthetic, it can be extremely useful to researchers conducting cybersecurity research. This is because each team is supposed to be representing a hypothetical company, which has to provide certain services to the outside world like any other company. Then similar to any company, they must deal with customer support, troubleshooting requests, new requests that the company must accomdate such as new services, new policies, verifying compliace, and so on. In addition, they have to deal with constant attacks of varying difficulty levels. Therefore, while this is semi-synthetic, we argue it is generally representative of organizations in that it contains operational benign data, attack flows, and defensive actions. Since it is already available, and it is semi-synthetic, privacy concerns are not an issue and this dataset can be downloaded through IMPACT's website.

## IV. Methodology

In our research, we are establishing ground truth of the NCCDC 2015 dataset and labeling the dataset. The 2015 dataset was selected because of the first-hand knowledge of the primary researcher being a member of one of the 2015 NCCDC blue teams. This adds distinctive subject matter expertise to the ground truth analysis and labeling process. The 2015 dataset contains PCAPS involving network activity for ten teams—each defending identical set-ups. The specific team flows being analyzed for ground truth are the those related to the primary researcher's experience in the competition.

Following previous literature, our plan to establish ground truth is as follows. The first phase consists of categorizing the traffic based on what service it corresponds with, such as an FTP service, web traffic, and so on. Such service level characterization provides insight into attacker tactics, techniques, and procedures as they relate to vulnerability exploitation strategy.

The second phase then looks at each of these services, and distinguishes what traffic is malicious and what traffic is not malicious. We leverage Snort to help us identify obvious malicious traffic, often of lesser sophistication. Traffic flagged by Snort will be reviewed to verify maliciousness.

In the third phase, we examine the remaining flows for attacks that were not detected by Snort. At this point, a more manual inspection of the data is needed, as the attacks are likely more sophisticated. For example, just because the source and destination IP addresses are known good IP addresses (i.e., internal addresses) does not make this traffic not malicious, as it could be lateral movement. Therefore, behavioral characteristics of the traffic are considered.

Once all traffic has been identified as malicious or benign, we label the traffic. We label it with standard attack metadata, as well as its 'location' in the Lockheed Martin Cyber Kill Chain. Subject matter experts are used in labeling. Results are then compared, and differing labels are discussed and conflicts resolved.

## V. Conclusion

In this paper, we have argued for the need for more sophisticated and realistic cybersecurity datasets. We introduced a dataset that contains more sophisticated and realistic attacks than widely cited datasets. Following this, we introduced our proposed methodology to help establish ground truth and label this dataset. We position our approach for and need of such a labeled dataset in our overall research aim of creating an autonomous intrusion detection system that specifically characterizes attack sophistication risk and correlates malicious network events across the Lockheed Martin Cyber Kill Chain. Our hope is that other researchers will be inspired by this dataset and our work, and will help create new novel techniques to improve the field of instrusion detection systems.

## References

[1] R. P. Lippmann *et al.*, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," in *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*, 2000, vol. 2, pp. 12–26.

[2] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Computer networks*, vol. 34, no. 4, pp. 579–595, 2000.

[3] S. Hettich and S. D. Bay, "The UCI KDD Archive." University of California, Department of Information and Computer Science, 1999.

[4] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[5] M. Canini, W. Li, and A. Moore, "GTVS: boosting the collection of application traffic ground truth," University of Cambridge Computer Labratory, Technical 748, Apr. 2009.

[6] S. Barnum, "Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX)," MITRE, 1.1, Feb. 2014.

[7] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," *Leading Issues in Information Warfare & Security Research*, vol. 1, p. 80, 2011.

[8] I. Ghafir and V. Prenosil, "Proposed Approach for Targeted Attacks Detection," in *Advanced Computer and Communication Engineering Technology*, vol. 362, H. A. Sulaiman, M. A. Othman, M. F. I. Othman, Y. A. Rahim, and N. C. Pee, Eds. Cham: Springer International Publishing, 2016, pp. 73–80.

[9] A. Lemay, J. Fernandez, and S. Knight, "Modelling physical impact of cyber attacks," in *Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES), 2014 Workshop on*, 2014, pp. 1–6.

[10] S. G. Batsell, N. S. Rao, and M. Shankar, "Distributed intrusion detection and attack containment for organizational cyber security," *Information on http://www. ioc. ornl. gov/projects/documents-/containment. pdf*, 2005.

[11] T. Sommestad, M. Ekstedt, and P. Johnson, "Cyber security risks assessment with bayesian defense graphs and architectural models," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, 2009, pp. 1–10.

[12] T. Sommestad, M. Ekstedt, and P. Johnson, "A probabilistic relational model for security risk analysis," *Computers & Security*, vol. 29, no. 6, pp. 659–679, 2010.

[13] N. L. Beebe and J. Guynes, "A model for predicting hacker behavior," *AMCIS 2006 Proceedings*, p. 409, 2006.

[14] Department of Homeland Security, "Impact Cyber Trust - DHS." 2016.