Towards a Reliable Intrusion Detection Benchmark Dataset

Iman Sharafaldin, Amirhossein Gharib, Arash Habibi Lashkari and Ali A. Ghorbani

Canadian Institute for Cybersecurity (CIC), UNB,
Fredericton, Canada
E-mail: {isharafa; agharib; a.habibi.l; ghorbani}@unb.ca

Received 15 March 2017; Accepted 15 June 2017;
Publication 7 July 2017

Abstract

The urgently growing number of security threats on Internet and intranet networks highly demands reliable security solutions. Among various options, Intrusion Detection (IDSs) and Intrusion Prevention Systems (IPSs) are used to defend network infrastructure by detecting and preventing attacks and malicious activities. The performance of a detection system is evaluated using benchmark datasets. There exist a number of datasets, such as DARPA98, KDD99, ISC2012, and ADFA13, that have been used by researchers to evaluate the performance of their intrusion detection and prevention approaches. However, not enough research has focused on the evaluation and assessment of the datasets themselves and there is no reliable dataset in this domain. In this paper, we present a comprehensive evaluation of the existing datasets using our proposed criteria, a design and evaluation framework for IDS and IPS datasets, and a dataset generation model to create a reliable IDS or IPS benchmark dataset.

Keywords: Intrusion Detection, Intrusion Prevention, IDS, IPS, Evaluation Framework, IDS dataset.

Journal of Software Networking, 177–200.
doi: 10.13052/jsn2445-9739.2017.009

1 Introduction

Intrusion detection has drawn the attention of many researchers whose aim are to address the ever-increasing issue of anomaly and unknown attacks. The adoption of IDS in real-world applications has been prevented by system complexity, so prior to deployment of IDS, a substantial amount of testing, evaluation, and tuning is needed. Evaluating IDS using labeled traffic enriched with intrusions and abnormal behavior is ideal but not always possible. Hence researchers normally resort to datasets that are often sub-optimal.

The ongoing change of network behaviors and patterns along with intrusion evolution, makes it necessary to move away from static and one-time datasets to dynamicly generated datasets. These most not only reflect the traffic compositions and intrusions, but also be modifiable, extensible, and reproducible [28]. However, the selection of a suitable dataset is a significant challenge itself since many such datasets are internal and cannot be shared due to privacy issues. Therefore, the available datasets are heavily anonymized and do not reflect the current real world trends. Because of the lack of certain statistical characteristics and the unavailability of these datasets a reliable dataset is yet to be realized [1, 28]. It is also necessary to mention that based on the malware evolution and continuous changes in attack strategies, benchmark datasets need to be updated periodically [28].

This paper is organized as follows. An overview of the datasets generated between 1998 and 2016 is presented in Section 2. The second section (Section 3) discusses the previous evaluation frameworks and provides details of our new framework. Section 3.1 presents an assessment and evaluation of the available datasets using the proposed framework. And finally Section 4 proposes an IDS dataset generating model which can cover all new criteria.

2 Available Datasets

In this section, some of the existing IDS datasets have been evaluated in order to identify the critical characteristics of a worthy dataset.

DARPA (Lincoln Laboratory 1998, 1999): This dataset was constructed for network security analysis purposes. Researchers criticized DARPA due to issues associated with the artificial injection of attacks and benign traffic. DARPA includes activities such as send and receive mail, browse websites, send and receive files using FTP, the use of telnet to log into remote computers and perform work, send and receive IRC messages, and monitor the router remotely using SNMP. It contains attacks like DOS, guess password, buffer

overflow, remote FTP, syn flood, Nmap, and rootkit. Unfortunately it does not represent real-world network traffic and contains irregularities such as the absence of false positives, and is outdated for the effective evaluation of IDSs on modern networks in terms of attack types and network infrastructure. Moreover, it lacks the actual attack data records [9, 10].

KDD'99 (University of California, Irvine 1998, 99): The KDD Cup 1999 dataset was created by processing the tcpdump portion of the 1998 DARPA dataset, which nonetheless suffers from the same issues. KDD99 includes more than twenty attacks such as neptune-dos, pod-dos, smurf-dos, buffer-overflow, rootkit, satan, teardrop, to name a few [3]. The network traffic records of normal and attack traffic are merged in a simulated environment resulting in a dataset with a large number of redundant records that are studded with data corruptions that leads to skewed testing results [11]. As a resolution to these shortcomings, NSL-KDD was created using the KDD dataset [11] to address some of KDD's shortcomings [9].

DEFCON (The Shmoo Group, 2000): Generated in 2000, DEFCON-8 dataset contains port scanning and buffer overflow attacks, whereas the DEFCON-10 dataset which was created in 2002 uses port scan and sweeps, bad packets, administrative privilege, and FTP by telnet protocol attacks. The traffic produced during the Capture the Flag (CTF) competition is different from the real world network traffic since it mainly consists of intrusive traffic as opposed to normal background traffic. This dataset is used to evaluate alert correlation techniques [21, 22].

CAIDA (Center of Applied Internet Data Analysis – 2002/2016): CAIDA consists of three different types of datasets: 1) CAIDA OC48, which includes different types of data observed on an OC48 link in San Jose and provided by CAIDA members, DARPA, NSF, DHS, Cisco; 2) CAIDA DDOS attack dataset, which includes one-hour DDoS attack traffic split of 5-minute pcap files; and 3) CAIDA Internet trace 2016, which is passive traffic traces from CAIDA's equinix-chicago monitor on High-speed Internet backbone. Most of CAIDAs datasets are very specific to particular events or attacks and are anonymized with their payload, protocol information, and destination. This dataset is not an effective benchmarking datasets due to a number of shortcomings, see [1, 16–19] for details.

LBNL (Lawrence Berkeley National Laboratory and ICSI – 2004/2005): LBNL's internal enterprise traces are full header network traffic recorded at a medium-sized site. This is a dataset without payload and suffers from a

heavy anonymization to remove any information which could identify an individual IP [20].

CDX (United States Military Academy 2009): The CDX dataset shows how the network warfare competitions can be utilized to generate modern day labeled dataset. In this dataset common attack tools namely Nikto, Nessus, and WebScarab have been used by attackers to carry out reconnaissance and attacks automatically. Benign traffic includes web, email, DNS lookups, and other required services. CDX can be used to test IDS alert rules but it suffers from lack of traffic diversity and volume [12].

Kyoto (Kyoto University – 2009): This dataset has been created using honeypots, so there is no process for manual labeling and anonymization, but it has limited view of the network traffic because only attacks directed at the honeypots can be observed. It has ten extra features such as IDS_Detection, Malware_Detection, and Ashula_Detection than the previous available datasets which are useful in NIDS analysis and evaluation. Since normal traffic is simulated repeatedly during the attacks and only produces DNS and mail traffic data, which does not reflect real world normal traffic, there are no false positives. False positives are important as they minimize the number of alerts [13–15].

Twente (University of Twente – 2009): To create this dataset, three services OpenSSH, Apache web server and Proftp using auth/ident on port 113 were installed to collect data from a honeypot network using netflow. Some side-effect traffic such as auth/ident, ICMP, and irc traffic which are not completely benign or malicious are generated. Moreover, it contains some unknown and uncorrelated alerts traffic. This dataset is labeled and is more realistic but the lack of volume and diversity of attacks is an obvious shortcoming. [5].

UMASS (University of Massachusetts – 2011): The dataset includes trace files which are network packets and some traces on wireless applications [26, 28]. It has been generated using a single TCP-based download request attack scenario. The dataset is not useful for testing IDS and IPS techniques due to the lack of variety of traffic and attacks [27].

ISCX2012 (University of New Brunswick – 2012): This dataset was generated by a dynamic approach and the authors present a good guideline for generating realistic and useful IDS evaluation datasets. Their approach consists of two parts namely Alpha and beta profile. The Alpha profile carries out various multistage attack scenarios to stream the anomalous segment of the dataset. While the Beta profile, which is the benign traffic generator, generates

realistic network traffic with background noise. Real traces are analyzed to create profiles to generate real traffic for HTTP, SMTP, SSH, IMAP, POP3, and FTP protocols. The dataset generated by this approach consists of network traces with full packet payloads and relevant profiles. However, it does not represent new network protocols since nearly 70% of todays network traffics are HTTPS and there are no HTTPS traces in this dataset. Moreover, the distribution of the simulated attacks is not based on real world statistics [1].

ADFA (University of New South Wales – 2013): To create the ADFA dataset, authors installed Apache, MySQL and Tikiwiki to offer a web service, Database server, remote access and FTP server. This dataset includes FTP and SSH password bruteforce, Java based Meterpreter, Linux Meter-preter payload and C100 Webshel attack vectors. It contains normal training and validating data and 10 attacks per vector [23]. In addition to lack of attack diversity and variety of attacks, the behavior of some attacks in this dataset are not well separated from the normal behavior [24, 25].

3 Proposed Framework

Over the years network attacks have evolved from a simple low-scale to a more sophisticated large-scale problem. During this period there has been considerable research and development into attack detection strategies, but only limited research in testing these techniques against realistic data. One of the key reasons was the legal and privacy issues associated with sharing captured data. As a result, the majority of the published works are evaluated by:

Replaying publicly available datasets: One obvious way to test and evaluate a detection system is to replay real attack traffic [2, 4].

Generating traffic: Artificial generation would seem to be the only practical way to generate both attack and benign traffic. Unfortunately, both hardware and software traffic generators are far from being ideal for simulating such attacks. Curl-loader is one of the open source tools to generate artificial traffic.

Testbed design strategies: An essential requirement for deploying synthetic traffic traces is to have an experimental setup and a traffic generation software. There are three commonly used testbed design strategies:

- Simulation: In this method, attackers, targets and network devices are all simulated [6]. NS-2 and Opnet are two examples of network simulators that are used.

- Emulation: It is a step forward in realism over simulation. Real machines are used as attackers and targets, and only the network topology is recreated in software [7].
- Direct physical representation: In this technique, the desired network topology is built by physically arranging a network of routers, switches, and computers [8].

The literature reviewed above highlights the shortcomings of various approaches used for building datasets. Although significant studies have been done on IDS dataset generation, little research has been conducted on the evaluation and assessment of IDS and IPS datasets. In proposing a new framework, we have taken into account the evaluation and assessments that have been done.

Scott *et al.* presented three major criteria in datasets: Redundancy, inherent unpredictability and complexity or multivariate dependencies [29]. Heide-mann and Papadopoulos used trace data in research to find common problems that cut across types of data and defined four aspects, namely privacy and anonymization, unavailable type of data such as local-observation or local-inference, developing new techniques, and moving target and coverage. They suggested that one of the most important aspects even when some data already exist is continued observation [31].

Ghorbani *et al.* discussed the IDS and evaluation criteria of these systems and believe that datasets are valuable assets in this domain. But since there are some issues in the creation of these datasets such as synthetic traffic generation or difficulties in collecting real attack scripts and victim software, all of them suffer from the fact that they are not good representatives of the real world traffic [30].

Nehinbe outlined an evaluation based on previous work regarding some aspects such as privacy issues, approval from owners, documentation problems, labeling, and anonymity [28]. Shiravi *et al.* defined evaluation criteria with six aspects: Realistic network, realistic traffic, labeled dataset, total interaction capture, complete capture, and diversity of attacks. They also assessed five previous datasets based on these aspects [1].

Given the shortcomings of the existing datasets, we believe that a comprehensive and wholesome framework for generating IDS/IPS benchmarking datasets is needed. The following section defines the features of such a framework. All in all, we define eleven features as follows:

1. **Complete Network Configuration**: Having a complete computer network is in fact the foundation of an offline dataset to represent the real

world. Several attacks have revealed their true faces only in a perfect network which has all equipment such as number of PCs, server(s), router, and firewall. So it is necessary to have a realistic configuration in the testbed to capture the real effects of attacks.

2. **Complete Traffic**: Traffic is a sequence of packets from a source (that canbe a host, router, or switch) to a destination (a host, a multicast group, or a broadcast domain). Based on the traffic generation techniques it is possible to have realistic, pseudo-realistic, or synthetic traffic in a dataset. The pseudo-realistic partially has the real world traffic, such as injecting network attacks into a benign dataset.
3. **Labeled Dataset**: While a dataset for evaluating different discovery mechanisms in this domain is important, tagging and labeling data are also important. If there are no correct labels, without a doubt, it is not possible to use a dataset and the results of the analysis are also not reliable. For example, in network datasets, only after converting pcaps to netflows it is possible to have reliable labels for flows. But, these are labeled datasets which do not clearly state the name and type of the attacks and only label them as benign or malicious. In other words, it is possible to have unlabeled, partially-labeled, and fully-labeled datasets.
4. **Complete Interaction**: For the correct interpretation of the results evaluation, one of the vital features is the amount of available information for anomalous behavior. Therefore, having all network interactions such as within or between internal LANs is one of the major requirements for a valuable dataset.
5. **Complete Capture**: Even in a complete traffic dataset, it is essential to capture all traffic, so as to be able to evaluate the proposed detection systems. It seems some of the datasets capture traffic partially and remove part of the traffic which is non-functional or not labeled. In order to have an accurate evaluation of IDS systems, it is neccessary for the dataset to complete capture. The removal of this traffic which have a significant role in the calculation of the false-positive rate of an IDS system, make the dataset weak.
6. **Available Protocols**: There are many different types of traffic, some of which are vital for testing an IDS system such as bursty traffic. Bursty traffic is an uneven pattern of data transmission and can cover some protocols such as HTTP and FTP. Interactive traffic includes sessions that consist of short request and response pairs such as applications involving real-time interaction with users (e.g., web browsing, online purchasing). In latency sensitive traffic the user has an expectation that

data will be delivered on time such as VOIP and Video conferencing. In Non-Real-time traffic such as news and mail traffics, timely delivery is not important. A complete dataset should have both normal and anomalous traffic.

7. **Attack Diversity**: In recent years, threats have expanded their scope into intricate scenarios such as application and app attacks. The types of attacks are changing and being updated on a daily basis. Therefore, having the ability to test and analyze IDS and IPS systems using these new attack and threat scenarios is one of the most important requirements that an off-line dataset should support. We categorized attacks into seven major groups based on the 2016 McAfee report, Browser-based, Brute force, DoS, Scan or enumeration, Backdoors, DNS, and other attacks (e.g., Heartbleed, Shellshock, and Apple SSL library bug).
8. **Anonymity**: The privacy compromising issues occur when both the IP and payload are available. So, most of the datasets remove their payload entirely which decreases the usefulness of the dataset especially for some detection mechanisms such as deep packet inspection (DPI).
9. **Heterogeneity**: In IDS domain, it is possible to have different sources for creating a dataset such as network traffic, operating systems logs, or network equipment logs. A homogeneous dataset with one type of source can be useful for analyzing a specific type of detection system while a heterogeneous dataset can be used for a complete test covering all aspects of the detection process.
10. **Feature set**: The main goal of providing a dataset is its usability for other researchers to test and analyze their proposed system. One of the main challenges is how to calculate and analyze the related features. It is possible to extract features from different types of data sources such as traffic or logs using feature extraction applications.
11. **Metadata**: Lack of proper documentation is one of the main issues in the available datasets in this area. Most of the datasets do not have documentation and if they do it is often incomplete. Insufficient information about the network configuration, operating systems for attacker and victim machines, attack scenarios, and other vital information can detract from the usability of a dataset for researchers.

Equation (1) is used for measuring the proposed framework. In this equation, W as a flexibility coefficient is the weight of each feature, which can be defined based on the organization request or type of the IDS system that has been selected for testing. Since we have eleven features in our framework,

we have to define eleven W for any scenario. V is the coefficient of each sub-factor that can be defined based on experiences or distribution of sub-factors in different scenarios. We have two features with sub-factors: Attacks and protocols. In these features V should be defined for each different sub-factor as well. Similarly, F is the appearance of a specific factor and sub-factor in the dataset that can be binary (0 or 1) or multi-valued.

$$\sum_{i=1}^{n} W_i \left(\sum_{j=1}^{m} V_j * F_j \right) \tag{1}$$

where n is the number of features and m is the number of coefficients for each factor. In the proposed framework, for two features "attacks" and "protocols" the value of m is 7 and 5 respectively but for the other features $m = 1$. To better understand the equation, the analysis of two datasets is shown in the next section with a specific value of W and V.

3.1 Evaluation of Current Datasets

Table 1 shows the weaknesses and strengths of eleven available datasets which have been listed and explained in Section 2 based on related documents and research and the proposed new dataset (Section 4); as is evident only the proposed dataset covers all criteria. Some of the features values are not shown because of lack of metadata and complete documentation. In order to apply the proposed framework to each dataset and calculate the score, it is necessary to define the W and V using a realistic scenario. Here we have selected two famous datasets KDD99 and KYOTO, to evaluate the framework.

As W is related to the organization or type of the IDS systems, we consider different values such as [0.05, 0.05, 0.1, 0.05, 0.05, 0.25, 0.25, 0.05, 0.05, 0.05, 0.05] in our scenario. Since protocol and attack factors have more values for us, we defined higher weights (0.25) for them. Further to define the V values for each factor, we have defined two different distributions for the attack and protocol. Based on the McAfee report [39] the distribution of seven attack categories are as follows: Browser (36%), Bruteforce (19%), DoS (16%), Scan (3%), DNS (3%), Backdoor (3%), Others (20%). It is necessary to mention that as SSL attacks are seasonal attacks and will not have a fixed value at all times, this attack was combined with "Others" in our distribution. This made the V values for attack factors to be as follows: [0.36, 0.19, 0.16, 0.03, 0.03, 0.03, 0.20].

Table 1 Comparing of available datasets based on evaluation framework

		DARPA	KDD'99	DEFCON	KAIDAs	LBNL	CDX	KYOTO	TWENTE	UMASS	ISCX2012	ADFA2013	Proposed Model
Network		Y	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y
Traffic		N	N	N	Y	Y	N	N	Y	N	N	Y	Y
Label		Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y
Interaction		Y	Y	Y	N	N	Y	Y	Y	N	Y	Y	Y
Capture		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Protocols	http	Y	Y	Y	–	Y	Y	Y	Y	Y	Y	Y	Y
	https	N	N	N	–	N	N	Y	N	N	N	N	Y
	ssh	Y	Y	Y	–	Y	Y	Y	Y	–	Y	Y	Y
	ftp	Y	Y	N	–	N	Y	Y	N	N	Y	Y	Y
	email	Y	Y	N	–	N	Y	Y	N	N	Y	Y	Y
Attacks	Browser	Y	Y	N	N	–	N	Y	N	N	Y	Y	Y
	Bruteforce	Y	Y	N	N	–	N	Y	Y	N	Y	Y	Y
	DoS	Y	Y	–	Y	–	Y	Y	N	–	Y	–	Y
	Scan	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Backdoor	N	N	Y	N	–	N	Y	N	N	N	N	Y
	DNS	N	N	N	Y	–	Y	Y	N	N	N	N	Y
	Others	Y	Y	Y	Y	–	–	Y	Y	Y	Y	Y	Y
Anonymity		N	N	–	Y	Y	–	N	–	–	N	–	Y
Hetrogenity		N	N	N	N	N	N	N	–	–	Y	–	Y
Feature Set		N	Y	N	N	N	N	Y	N	N	N	N	Y
Metadata		Y	Y	N	Y	N	N	Y	Y	N	Y	Y	Y

One of the other shortcomings of the available datasets is the distribution of the protocols. The rapid growth of the Internet has changed the protocols distributions deistically hence finding a valid document will be very difficult. As a way forward, we observed the traffic of our research center for one month to find this distribution. For a group of six protocols the usage percentage and distribution was http (10%), https (74%), ssh (2%), ftp (6%), email (1%), and other (7%). So, the V vaules for protocol factor will be [0.1, 0.74, 0.04, 0.08, 0.04].

Table 2 shows the value of the KDD and KYOTO datasets in this scenario which are 0.56 and 0.85, respectively. Figure 1 shows the comparison between KDD and KYOTO datasets based on the binary values from Table 1 and the scores from Table 2.

Table 2 KDD99 and KYOTO datasets scores

Dataset	Calculation	Score
KDD	0.05*1 + 0.05*0 + 0.1*1 + 0.05*1 + 0.05*1+ 0.25 * (0.1 + 0.0 + 0.04 + 0.08 + 0.04) + 0.25 * (0.0 + 0.19 + 0.16 + 0.03 + 0.0 + 0.0 + 0.2) + 0.05*0 + 0.05*0 + 0.05*1 + 0.05*1	0.56
KYOTO	0.05*1 + 0.05*0 + 0.1*1 + 0.05*1 + 0.05*1 + 0.25 * (0.1 + 0.74 + 0.04 + 0.08 + 0.04) + 0.25 * (0.36 + 0.19 + 0.16 + 0.03 + 0.03 + 0.03 + 0.2) + 0.05*0 + 0.05*0 + 0.05*1 + 0.05*1	0.85

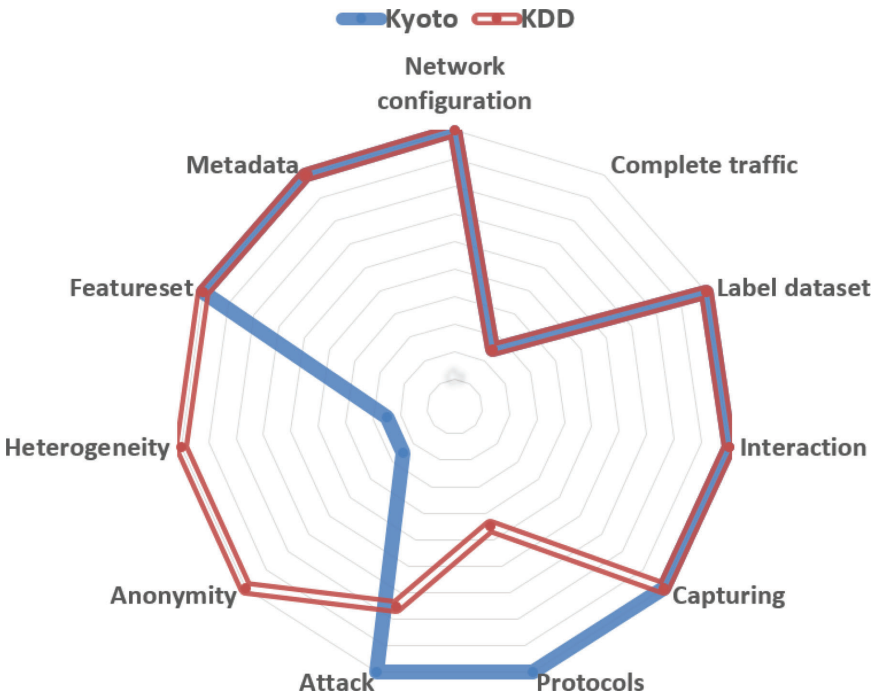

Figure 1 KDD99 and KYOTO datasets evaluation.

4 Generating Reliable Dataset

In this section, we present a systematic approach to generate a realistic IDS dataset. It consist of two components namely `B-Profile` and `M-Profile`.

The `B-Profile` is responsible for profiling the abstract behaviour of human interactions and generate a naturalistic benign background traffic. The `M-Profile` is used to describe and perform attack scenarios. Profiles can be

applied to a diverse range of network protocols with different topology, because they represent the abstract properties of human and attack behavior. To have a complete representation of real world network, it is crucial to include gateway, router, switch, servers and PCs in the implemented network.

B-Profile: To produce benign background traffic, `B-Profile` is designed to extract the abstract behaviour of a group of human users. It tries to encapsulate network events produced by users with machine learning and statistical analysis techniques. The encapsulated features are distributions of packet sizes of a protocol, number of packets per flow, certain patterns in the payload, size of payload, and request time distribution of protocols. Once B-Profiles are derived from users, an agent or human operator can use them to generate realistic benign events and simultaneously perform M-Profile on the network. Organizations and researchers can use this approach to easily generate realistic data sets; therefore, there is no need to anonymize data sets. As shown in Figure 2, benign profile models are created in two steps:

- **Individual Profiling:** The most popular protocols in network traffic are HTTP, HTTPS, FTP, SSH, and email protocols which a rich dataset should contains events from all of them. There are several ways to capture user's activities such as Man In The Middle (MITM), network sniffing, browser and email histories. The input for the first step is the users' behaviours in term of mentioned protocols. Network activities of each user are recorded daily (day and protocol) and a histogram of events with 48 bars (every 30 minutes) is calculated. Figure 3 shows the individual profile of a user for one day.
- **Clustering:** In the clustering step, individual user profiles are analyzed against other users to create clusters of users with similar behaviour and distribution. In real world scenario, it is impossible to determine the exact number of groups since each group represents one particular abstract

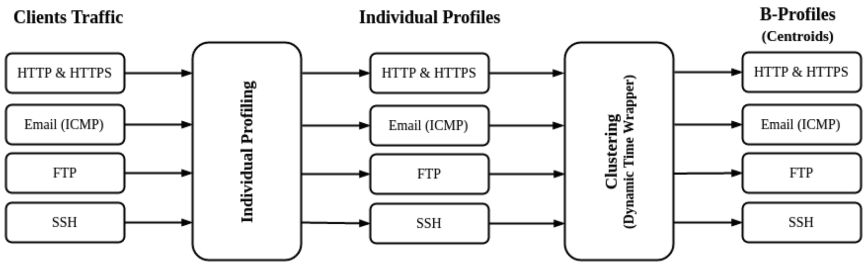

Figure 2 Benign profiling design.

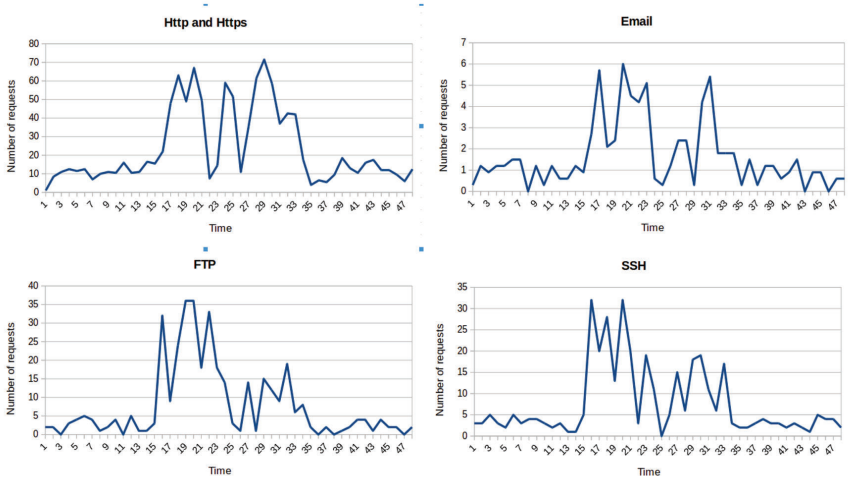

Figure 3 Individual profile of a user for one day.

behaviour. Therefore, to have a flexible clustering, XMeans clustering algorithm [36] is preferred to other algorithms such as KMeans since it can automatically learn the number of clusters.

One of the influential factors in clustering is the distance function. In some applications choice of the distance function is more important than the clustering algorithm itself [40]. Since individual profiles are time series data, classic distance function such as Jaccard, Euclidean, and Cosine show poor performance in clustering them. We have used Dynamic Time Warping (DTW) [37] for XMeans clustering. DTW is an algorithm for measuring similarity between two given time-dependent sequences and provides a better performance than classic distance functions since the individual profile is a time-dependent profile.

DTW is an extension of Euclidean distance that employ a local (non-linear) alignment. Given two time series $\overrightarrow{a} = (a_1, \ldots, a_m)$ and $\overrightarrow{b} = (b_1, \ldots, b_m)$, it creates an *m*-by-*m* matrix *M* with the Euclidean Distance (ED) between any two points of $\overrightarrow{a}$ and $\overrightarrow{b}$. A warping path $W = (w_1, \ldots, w_n)$, where $n >= m$, is a contiguous group of matrix items which represent a mapping between $\overrightarrow{a}$ and $\overrightarrow{b}$ [38]:

$$DTW(\vec{a}, \vec{b}) = \min \sqrt{\sum_{i=1}^{n} w_i} \tag{2}$$

The path is computable by means of dynamic programming with the following formula:

$$\alpha(i, j) = ED(i, j) + \min\{\alpha(i-1, j-1), \alpha(i-1, j), \alpha(i, j-1)\} \tag{3}$$

Finally, the centroid of each cluster is calculated and considered to be the abstract behaviour of users in that cluster. To generate traffic from `B-Profiles` at the start of the day, an agent will randomly select a `B-Profile` from the available collection of profiles to mimic human behavior. To keep the requests as realistic as possible, a slightly modified web-crawling mechanism is devised to demonstrate the browsing behaviors of users for HTTP and HTTPS. The Agent function is a multi-threaded Python script that can execute requests in parallel. The same procedure can be used for the other protocols as well.

M-Profile: This section tries to define attack scenarios in an unambiguous manner. In the simplest case, users can interpret these profiles and subsequently apply them. In an ideal case, autonomous agents along with compilers would be employed to interpret and execute these scenarios. Six different attack scenarios are considered, which can cover all of the attack categories listed on [39]:

- **Infiltration of the network from inside:** In this scenario a vulnerable document viewer should be exploited. First victim receives a malicious document through the email. After successful exploitation, a backdoor will be executed on the victims computer and then the backdoor start to scan the internal network for other vulnerable boxes, and exploit them if it is possible.
- **Unsuccessful infiltration of the network from inside:** Attacks are not always successful and this issue should be reflected in the dataset to make it more realistic, The following scenario is defined:
  - Send an email which contains a malicious URL to the victims on behalf of admin@[...] (using email spoofing)
  - Exploit a browser vulnerability and get a reverse shell (as a limited user)
  - Unsuccessfully try to do privilege escalation to elevate current user privilege

- Acquire browser cookies of users for different sites (as we could not acquire administrator user in the previous step)
- Get access to victim's accounts through collected cookies

Denial of Service (DoS): As DoS is one the prominent network attacks, this attack also should be available in any dataset. In this scenario, HTTP DoS is utilized; there are some tools available which can make web servers completely inaccessible using a single machine. These tools start by making a full TCP connection to the remote server. The tool holds the connection open by sending valid, incomplete HTTP requests to the server at regular intervals to keep the sockets from closing. Since any web server has a finite ability to serve connections, it will only be a matter of time before all sockets are used up and no other connection can be made.

- **Collection of web application attacks:** In this scenario a vulnerable web application is developed. In the first step, victim website is scanned through a web application vulnerability scanner, then the attacker can conduct different types of web attacks on the vulnerable website (including SQL injection, command injection, unrestricted file upload).
- **Brute force attacks:** Brute force attacks are common against networks as they tend to break into accounts with weak username and password combinations. The final scenario has been designed with the goal of acquiring different services accounts by running a dictionary brute force attack against the main server (at least two major services should be chosen).
- **Recent attacks:** There are some attacks based on famous vulnerabilities that can be conducted during a specific amount of time. These are extraordinary vulnerabilities which sometimes affects millions of servers or victims, and normally it takes months to patch all vulnerable computers around the world. Some of the most famous vulnerabilities in recent years are Heartbleed, Shellshock and Apple SSL library bug.

4.1 Analysis and Evaluation

According to the proposed evaluation framewrok a complete dataset should cover eleven criteria. As the last row on the Table 1 shows, the new dataset generating model comprises them all. To cover the `complete network configuration` criteria, a network which includes gateway, router, switch, servers and PCs has been implemented.

As presented in the previous Section 4, the proposed dataset generator is composed of two parts: M-Profile and B-Profile. M-Profile generates malicious traffic by performing a variety of real world attack scenarios. On the other hand, B-Profile can easily generate real world benign traffic by profiling and executing user behavior. Hence, these two profiles guarantee the generation of a complete traffic dataset.

In accordance with the mentioned network topology, the dataset generator provides complete benign network interaction. Moreover for the malicious traffic, M-Profile provides complete network interaction as it considers both infiltration from inside and outside network attacks. Further, as the existence of the dataset's complete capture is essential in evaluating IDS systems, our dataset generator can fulfill this criterion by considering whole network traffic.

In respect of modern networks, our dataset generator is designed in a way to generate traffic which contains popular network protocols including HTTP, HTTPS, SSH, FTP, SMTP, IMAP and POP3. Consequently, the generated traffic reflects real world network protocols that are necessary for evaluating IDSs.

Moreover, for the malicious part of the dataset, we record the complete traffic and interaction data when we conduct each attack. According to Section 4, a complete range of attacks has been proposed on different scenarios (M-Profile) which is supporting the attack diversity criterion. Of course, having all generated network traffic and system logs along with resource usage, such as memory and CPU during experiments indicates supporting the heterogeneity criterion.

For labeling and tagging the dataset, we suggest two layer labeling. Firstly, we label the data based on the benign and attack days. Secondly, for each group of features, such as network traffic and resource usage, we tag each record based on the name of the specific attack. The main advantage of this type of labeling is having a reliable tag for each traffic flow record and related set of features which is vital for future analysis and characterization.

On the other hand, saving trace files and extracted feature set is one of the emerging requests of researchers in this domain. In this dataset, we also need to present a set of features in network traffic such as protocol, source bytes, destination bytes, sending packets, receiving packets, and flow size, as well as resource usage features, such as running service rate, processor usage, memory usage and service error rate.

Finally, to have sufficient information about the dataset, it is necessary to cover four sections in the metadata and documentation. It should explain

the network topology and give complete information about the network environment such as the number of machines, operating systems, and the common installed applications. In the benign section, it is essential to know the number of users and explain the type of activities and related application that agents run on each machine. The malicious section should have the types of intrusions and attacks which are simulated. As the last part, presenting the labeling method and having a list of extracting features with clear definitions is crucial.

5 Conclusion

In this paper, we have studied the existing IDS datasets in order to find the characteristics of a reliable dataset. We also presented a new framework to evaluate datasets with eleven criteria: attack diversity, anonymity, available protocols, complete capture, complete interaction, complete network configuration, complete traffic, feature set, heterogeneity, labeled dataset, and metadata. The proposed framework considers organization policy and conditions using a coefficient, W, which can be defined separately for each criterion. Furthermore, we proposed an ideal IDS dataset generating model that can cover all eleven criteria. In the future, we plan to generate a new dataset based on the proposed model and make it publicly available.

References

[1] Shiravi, A., Shiravi, H., Tavallaee, M., and Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Sec*. 31, 357–374.

[2] Ejaz, A., George, M., Alan, T., and Sajal, B. (2010). *Use of IP Addresses for High Rate Flooding Attack Detection*. Berlin: Springer, 124–135.

[3] KDD Cup (1999). *University of California, Irvine (UCI)*. Available at: http://kdd.ics.uci.edu/databases/kddcup99/

[4] Buchanan, B., Flandrin, F., Macfarlane, R., and Graves, J. (2011). A *Methodology to Evaluate Rate-Based Intrusion Prevention System against Distributed Denial-of-Service DDoS*. Glasgow: Cyberforensics University of Strathclyde.

[5] Sperotto, Anna and Sadre, Ramin and Vliet, Frank and Pras, Aiko. "A labeled data set for flow-based intrusion detection," in *Proceedings of the 9th IEEE International Workshop on IP Operations and Management IPOM09*, Venice, 39–50.

[6] Mirkovic, J., and Fahmy, S., and Reiher, P. and Thomas, R. K. (2009). "How to test dos defenses," in *Proceedings of the Conference For Homeland Security, CATCH '09, Cybersecurity Applications Technology*, 103–117.

[7] Rajab, M. A., Zarfoss, J., Monrose, F., Terzis, A. (2007). "My botnet is bigger than yours maybe, better than yours," in *Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets* (Berkeley, CA: USENIX Association), 5–10.

[8] Yu, Jaehak and Kang, Hyunjoong and Park, Daeheon and Bang, Hyo-Chan and Kang, Do Wook. An In-depth Analysis on Traffic Flooding Attacks Detection and System Using Data Mining Techniques, Journal System Architect, 59, 10, 1005–1012, Elsevier North-Holland, New York, NY, USA, 2013.

[9] McHugh, J. (2000). Testing Intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. Inform. Syst. Sec. 3, 262–294.

[10] Brown, C., Cowperthwaite, A., Hijazi, A., Somayaji, A. (2009). "Analysis of the 1999 DARPA/lincoln laboratory IDS evaluation data with NetADHICT," in *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Piscataway, NJ, 1–7.

[11] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Piscataway, NJ, 1–6.

[12] Benjamin Sangster, T. J., OConnor, Cook, T., Fanelli, R., Dean, E., Adams, W. J., Morrell, C., et al. (2009). *Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets* (Berkeley, CA: Usenix, The Advanced Computing System Association), 2009.

[13] Song, J., Takakura, H., Okabe, Y., Eto, M., and Inoue, D., and Nakao, K. (2011). "Statistical analysis of honeypot data and building of kyoto 2006+ dataset for NIDS evaluation," in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security* (New York, NY: ACM), 29–36.

[14] Sato, M., Yamaki, H., and Takakura, H. (2012). "Unknown attacks detection using feature extraction from anomaly-based ids alerts," in *Proceedings of the IEEE/IPSJ 12th International Symposium on Applications and the Internet (SAINT)*, Piscataway, NJ, 273–277.

[15] Chitrakar, R., and Huang, C. (2012). "Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and naive bayes classification," in *Proceedings of the 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, Piscataway, NJ, 1–5.

[16] Proebstel, E. P. (2008). *Characterizing and Improving Distributed Network-based Intrusion Detection Systems(NIDS): Timestamp Synchronization and Sampled Traffic*. (Davis, CA: University of California DAVIS).

[17] CAIDA (2002). *CAIDA data set OC48 Link A*. Available at: https://www.caida.org/data/passive/passive oc48 dataset.xml

[18] CAIDA (2007). *CAIDA DDoS Attack Dataset*. Available at: https://www.caida.org/data/passive/ddos-20070804 dataset.xml

[19] CAIDA (2016). *CAIDA Anonymized Internet Traces 2016 Dataset*, Available at: https://www.caida.org/data/passive/passive 2016 dataset. xml

[20] Nechaev, B., Allman, M., Paxson, V., and Gurtov, A. (2004). Lawrence Berkeley National Laboratory (LBNL)/ICSI Enterprise Tracing Project. (Berkeley, CA: LBNL/ICSI).

[21] DEFCON 8, 10 and 11, The Shmoo Grouphttp://cctf.shmoo.com, 2000.

[22] Nehinbe, J. O., and Weerasinghe, D. (2010). "A Simple Method for Improving Intrusion Detections in Corporate Networks," in *Information Security and Digital Forensics First International Conference ISDF* (Berlin: Springer), 111–122.

[23] Creech, G., and Hu, J. (2013). "Generation of a new IDS test dataset: time to retire the KDD collection," in *Proceedings of the 2013 IEEE Wireless Communications and Networking Conference (WCNC)*, New York NY, 4487–4492.

[24] Xie, M., and Hu, J. (2013). "Image and Signal Processing (CISP), 2013 6th International Congress on," in *Proceedings of the Evaluating host-based anomaly detection systems: A preliminary analysis of ADFALD*, Vol. 03 (Berlin: Springer), 1711–1716.

[25] Xie, M., Hu, J., and Slay, J. (2014). "Evaluating host-based anomaly detection systems: application of the one-class SVM algorithm to ADFA-LD," in *Proceedings of the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Xiamen, 978–982.

[26] U Mass Trace Repository (2011). *Optimistic TCP ACKing, University of Massachusetts Amherst*, Availabel at: http://traces.cs.umass.edu

[27] Prusty, S., Levine, B. N., and Liberatore, M. (2011). "Forensic investigation of the oneswarm anonymous filesharing system," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)* (New York, NY: ACM).

[28] Nehinbe, J. O. (2011). "A critical evaluation of datasets for investigating IDSs and IPSs researches," in *Proceedings of the IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*, 92–97, New York, NY.

[29] Scott, P. D., and Wilkins, E. (1999). Evaluating data mining procedures: techniques for generating artificial data sets. *Inform. Softw. Technol*. 41, 579–587, 1999.

[30] Ali, G., Wei, L., and Mahbod, T. (2010). *Network Intrusion Detection and Prevention: Concepts and Techniques*. New York, NY: Springer.

[31] Heidemann, J., and Papdopoulos, C. (2009). "Uses and challenges for network datasets," in *Proceedings of the Cybersecurity Applications Technology Conference For Homeland Security, CATCH'09*, Darlinghurst, NSW, 73–82.

[32] AsSadhanl, B. A. (2009). *Network Traffic Analysis Through Statistical Signal Processing Methods*. Pittsburgh, PA: Carnegie Mellon University.

[33] (2014). *Heartbleed, Codenomicon*, Available at: https://www.heartbleed.com

[34] (2016). *Shell Shock*. Available at: https://en.wikipedia.org/wiki/Shellshock

[35] (2016). *Critical Https Bug May open 25000 ios-apps to Eavesdropping Attacks*. Available at: http://arstechnica.com/security/2015/04/critical-https-bug-may-open-25000-ios-appsto-eavesdropping-attacks

[36] Pelleg, D., and Moore, A. W. (2000). "X-means: extending K-means with efficient estimation of the number of clusters." in *Proceedings of the ICML Seventeenth International Conference on Machine Learning*, Vol. 1 (San Francisco, CA: ACM).

[37] Mller, M. (2007). "Dynamic time warping," in *Information Retrieval for Music and Motion* (Berlin: Springer), 69–84.

[38] Paparrizos, J., and Gravano, L. (2015). "k-shape: efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (New York, NY: ACM).

[39] (2016). *McAfee Threat Report*. Available at: http://www.mcafee.com/ca/resources/reports/rp-quarterly-threatsmar-2016.pdf

[40] Batista, G. E. A. P. A. (2014). CID: an efficient complexity-invariant distance for time series. *Data Min. Knowl. Dis.* 28.3, 634–669.

Biographies

I. Sharafaldin received the M.Eng. degree in software engineering from the university of Tehran, Iran. He is currently working toward the Ph.D. degree in the Canadian Institute for Cybersecurity (CIC), University of New Brunswick, Canada. His research interests include network security, visual analytics, network traffic generation, botnet detection.

A. Gharib is a master student and research assistant at the Canadian Institute for Cybersecurity (CIC), University of New Brunswick, Canada. He obtained his B.Eng. from Ferdowsi University of Mashhad, Iran. He is currently involved in some cybersecurity projects such as IDS dataset generation, ransomware detection and security information & event management (SIEM) enhancement. His research interests include network security, malware analysis, and big security data analytic.

A. H. Lashkari is a research associate at Canadian Institute for Cybersecurity (CIC) on the Faculty of Computer Science, University of New Brunswick. He has more than 21 years of academic and industry experience. He has been awarded 3 gold medals as well as 12 silver and bronze medals in international competitions around the world. In addition, he is the author of 12 books in English and Persian on topics including cryptography, network security, and mobile communication as well as over 80 journals and conference papers concerning various aspects of computer security. His current research focuses on cybersecurity, big security data analysis, Internet Traffic Analysis and the detection of malware and attacks.

A. A. Ghorbani has held a variety of positions in academia for the past 35 years and is currently the Canada Research Chair (Tier 1) in Cybersecurity, the Dean of the Faculty of Computer Science (since 2008), and the Director of the Canadian Institute for Cybersecurity. He is the co-inventor on 3 awarded patents in the area of Network Security and Web Intelligence and has published over 200 peer-reviewed articles during his career. He has supervised over 160 research associates, postdoctoral fellows, graduate and undergraduate students during his career. His book, Intrusion Detection and Prevention Systems: Concepts and Techniques, was published by Springer

in October 2010. In 2007, Dr. Ghorbani received the University of New Brunswick's Research Scholar Award. Since 2010, he has obtained more than $10M to fund 6 large multi-project research initiatives. Dr. Ghorbani has developed a number of technologies that have been adopted by high-tech companies. He co-founded two startups, Sentrant and EyesOver in 2013 and 2015. Dr. Ghorbani is the co-Editor-In-Chief of Computational Intelligence Journal. He was twice one of the three finalists for the Special Recognition Award at the 2013 and 2016 New Brunswick KIRA award for the knowledge industry.