Report of Existing Intrusion Detection Datasets

Rui Shu
North Carolina State University
Raleigh, NC
rshu@ncsu.edu

Tim Menzies
North Carolina State University
Raleigh, NC
timm@ieee.org

ABSTRACT

We collect a bunch of datasets used in security related papers. The aim of this work is to figure out how these datasets are used (i.e., dependent variables, independent variables). We care about the data pre-processing, research goals (range), what are the evaluation measurement, scenarios., percentage of abnormal data/records in total data/records, etc.

KEYWORDS

Intrusion Detection

ACM Reference Format:
Rui Shu and Tim Menzies. 2018. Report of Existing Intrusion Detection Datasets. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In network intrusion detection (IDS), anomaly-based approaches in particular suffer from accurate evaluation, comparison, and deployment which originates from the scarcity of adequate datasets. Many such datasets are internal and cannot be shared due to privacy issues, others are heavily anonymized and do not reflect current trends, or they lack certain statistical characteristics. These deficiencies are primarily the reasons why a perfect dataset is yet to exist. Thus, researchers must resort to datasets which they can obtain that are often suboptimal.

What traits contribute to a **good security dataset**? [38]

(1) Realistic. A good dataset should ideally reflects the real effects attacks and corresponding response of the victims. So, it is necessary for both the attackers and the victims to behave as realistically as possible. Artificial adjustments or post-capture trace insertion that negatively affect the raw data and introduce inconsistency to the dataset should be avoided.
(2) Labeled Data. A labeled data allows for the distinction between normal and anomalous activities, which eliminating the impractical process of manual labeling.
(3) Large amount. A dataset with larger amount is far more important for us to detect the abnormal activities than a small amount of dataset.
(4) Complete Capture. Privacy concerns usually be the obstacle for researchers to share their complete datasets. Limited datasets or heavily anomayzed datasets usually removed some features or traces that results in decreased utility.
(5) Diverse scenarios.

2 EXISTING IDS DATASET

2.1 ADFA Linux Dataset (ADFA-LD12)

Cited papers: [11] [10] [45]
Description: This dataset is generated based on Ubuntu Linux version 11.04, which runs a Linux web server and services with known vulnerabilities. It provides some services such as file sharing, database services, remote access and web server. This dataset recorded the traces under attacks as described in the following table.

Payload/Effect	Attack Vector
Password bruteforce	FTP by Hydra
Password bruteforce	SSH by Hydra
Add new superuser	Client side poisoned executable
Java Base Meterpreter	Tiki Wiki vulnerability exploit
Linux Meterpreter payload	Client side poisoned executable
C100 Webshell	PHP Remote File Inclusion vulnerability

Table 1: Attack Structure

Research Goals: ADFA-LD12 is designed for anomaly based systems, not for signature recognition IDS. Compared with old datasets such as KDD 98 and KDD 99, this dataset is much more representative of current attacks, and forms a realistic and relevant metric for IDS performance metrics. Authors in the papers proposed a new host-based anomaly detection method using discontiguous system call pattern on the ADFA dataset in an attempt to increase detection rate while reducing false alarm rates.
Data Structure: This dataset contains three different data groups, and each group contains raw system call traces. Each training or validation data trace was collected during normal operation of the host. Traces are generated using the auditd Unix program and then filtered by size. The training data has a size with range 300 Bytes and 6KB, while the validation data keeps a size between 300 Bytes and 10KB. However, we could not get more details to infer the attributes from the description of the paper.

Evaluation Measurements: The validity of the data set was examined by evaluating the performance of several IDS algorithms,

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Data Type	Trace Count
Normal Training Data	833 Traces
Normal Validation Data	4373 Traces
Attack Data	100 Attacks per Vector

Table 2: Dataset Structure

i.e., hidden Markov models, the STIDE approach, K-Means clustering, and the K-Nearest Neighbour algorithm, and proposed semantic based methods. The result shows that the proposed method achieves a higher detection rate with a lower false alarm rate when compared with other algorithms.

2.2 CDX 2009

Cited papers: [16] [35] [8]
Description: This dataset contains data captured by NSA, data captured outside of the West Point network border and snort intrusion prevension log. 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.
Research Goals: To create a network based detection system for online defense against zero-day buffer overflow attacks in the production environment [16].
Data Structure: This dataset is a labeled dataset, and TCP dump traces includes all simulated communications and snort logs include information about the occurances of intrusions.
Evaluation Measurements:

2.3 KDD CUP 98&99 Dataset

Cited papers: [3] [41]
Description: This dataset is an updated version of the DARPA98, by processing the tcpdump portion. It contains different attacks such as Neptune-DoS, pod-DoS, Smurf-DoS, and buffer-overflow. The KDD experiments are performed on a Solaris-based system to collect a wide range of data. System calls are generated by processing the BSM audit data. **Research Goals:** This dataset is usually used to evaluate new intrusion detection methods based on analyzing network traffic.
Data Structure: The training dataset consists of approximately 4.9 million single connection vectors, each labeled as either normal or attack, containing 41 features per connection record.
Evaluation Measurements:

2.4 DARPA Intrusion Detection Data Sets

Cited papers: [7] [26]
Description: The datasets contained labeled data generated by simulating network traffic for a medium size U.S. Air Force base. 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 dataset was constructed for network security analysis and exposed the issues associated with the artificial injection of attacks and benign traffic. This dataset includes email, browsing, FTP, Telnet, IRC, and SNMP activities. It contains attacks such as DoS, Guess password, Buffer overflow, remote FTP, Syn flood, Nmap, and Rootkit.
Research Goals: The labeled DARPA datasets of 1998 and 1999 were the security community benchmark for testing intrusion. detection systems.
Data Structure:
Evaluation Measurements:

2.5 Intrusion detection evaluation dataset (ISCXIDS2012)

Cited papers: [38]
Description: The UNB ISCX IDS 2012 dataset consists of labeled network traces, including full packet payloads in pcap format, which along with the relevant profiles are publicly available for researchers.
Research Goals: The UNB ISCX 2012 Intrusion Detection Evaluation Data Set pocesses the following characteristics: Realistic network and traffic, labeled dataset, total interaction capture, complete capture, diverse intrusion scenarios.
Data Structure: The UNB ISCX 2012 intrusion detection evaluation dataset consists of the following 7 days of network activity (normal and malicious). The UNB ISCX IDS 2012 dataset consists of labeled network traces, including full packet payloads in pcap format, which along with the relevant profiles.
Evaluation Measurements:

2.6 Intrusion Detection Evaluation Dataset (CICIDS2017)

Cited papers: [37]
Description: CICIDS2017 dataset contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source and destination IPs, source and destination ports, protocols and attack (CSV files).
Research Goals:
Data Structure: The CICIDS2017 dataset consists of labeled network flows, including full packet payloads in pcap format, the corresponding profiles and the labeled flows and CSV files for machine and deep learning purpose are publicly available for researchers. The data capturing period started at 9 a.m., Monday, July 3, 2017 and ended at 5 p.m. on Friday July 7, 2017, for a total of 5 days. Monday is the normal day and only includes the benign traffic. The implemented attacks include Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS. They have been executed both morning and afternoon on Tuesday, Wednesday, Thursday and Friday.
Evaluation Measurements:

2.7 NSL-KDD dataset

Cited papers: [41] [6]
Description: NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set. The number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.
Research Goals:
Data Structure: The NSL-KDD dataset consists of two parts: (i) KDDTrain+ and (ii) KDDTest+. The KDDTrain+ part of the NSL-KDD dataset is used to train a detection method or system to detect network intrusions. It contains four classes of attacks and a normal class dataset. The KDDTest+ part of NSLKDD dataset is used for testing a detection method or a system when it is evaluated for performance. It also contains the same classes of traffic present in the training set
Evaluation Measurements:

2.8 VPN-nonVPN dataset (ISCXVPN2016)

Cited papers: [12]
Description: This dataset captures a regular session and a session over VPN, therefore we have a total of 14 traffic categories: VOIP, VPN-VOIP, P2P, VPN-P2P, etc.
Research Goals:
Data Structure: The traffic was captured using Wireshark and tcpdump, generating a total amount of 28GB of data. For the VPN, they used an external VPN service provider and connected to it using OpenVPN (UDP mode). To generate SFTP and FTPS traffic they also used an external service provider and Filezilla as a client.
Evaluation Measurements:

2.9 Botnet dataset

Cited papers: [4]
Description: To merge these data traces in one unified data set they employed so called overlay methodology, one of the most popular methods for creating synthetic datasets. Malicious data is usually captured by honeypots or through infecting computers with a given bot binary in a controlled environment. Botnet traces can be merged with benign data by mapping malicious data to either machines existing in the home network or machines outside of the current network. Considering the wide range of IP addresses in the traces, we mapped botnet IPs to the hosts outside of the current network using Bit-Twist packet generator. Malicious and benign traffic were then replayed using TCPReplay and captured by TCPdump as a single dataset.
Research Goals:
Data Structure: The dataset is divided into training and test datasets that included 7 and 16 types of botnets, respectively. The training dataset is 5.3 GB in size of which 43.92% is malicious and the reminder contains normal flows. Test dataset is 8.5 GB of which 44.97% is malicious flows.
Evaluation Measurements:

2.10 Android validation dataset

Cited papers: [14]
Description: For the validation dataset 72 unique Android apps were selected from different sources. They manually selected one sample from each family of the Android Malware Genome Project dataset, eight samples from the Android Malware Virus Share package, nine samples from the Virus Total and five samples from the official Google Play market.
Research Goals:
Data Structure: This data set which consist in 72 original apps from different origins, and 10 transformations on each app, resulting in 792 apps with different relationships between them. The transformations performed featured the following operations: insert junk code, insert junk files, replace icons, replace files, different aligns, replace strings and more.
Evaluation Measurements:

2.11 Android Botnet dataset

Cited papers: [18]
Description: To give a comprehensive evaluation of Android botnets, they gathered a large collection of Android botnet samples representing 14 botnet families. Their accumulated dataset combines some botnet samples from the Android Genome Malware project, malware security blog, VirusTotal and samples provided by well-known anti-malware vendor.
Research Goals:
Data Structure: Overall, their dataset includes 1929 samples spawning a period of 2010 (the first appearance of Android botnet) to 2014.
Evaluation Measurements:

2.12 Tor-nonTor dataset (ISCXTor2016)

Cited papers: [21]
Description: To be sure about the quantity and diversity of this dataset in CIC, they defined a set of tasks to generate a representative dataset of real-world traffic. They created three users for the browser traffic collection and two users for the communication parts such as chat, mail, FTP, p2p, etc. For the non-Tor traffic they used previous benign traffic from VPN project and for the Tor traffic they used 7 traffic categories.
Research Goals:
Data Structure: The traffic was captured using Wireshark and tcpdump, generating a total of 22GB of data. They captured the outgoing traffic at the workstation and the gateway simultaneously, collecting a set of pairs of .pcap files: one regular traffic pcap (workstation) and one Tor traffic pcap (gateway) file. They labelled the captured traffic in two steps. First, we processed the .pcap files captured at the workstation: we extracted the flows, and they confirmed that the majority of traffic flows were generated by application X (Skype, ftps, etc.), the object of the traffic capture. Then, they labelled all flows from the Tor .pcap file as X.

Evaluation Measurements:

2.13 CIC DoS dataset

Cited papers: [17]
Description: They have set up a testbed environment with a victim web server running Apache Linux v.2.2.22, PHP5 and Drupal v.7 as a content management system. The attacks were selected to represent the most common types of application layer DoS.
Research Goals:
Data Structure: Generated application layer DoS attacks were intermixed with the attack-free traces from the ISCX-IDS dataset. We produced 4 types of attacks with different tools, obtaining 8 different application layer DoS attack traces. These attacks were directed towards 10 web servers in ISCX data set that have the top highest number of connections. The resulting set contains 24 h of network traffic with total size of 4.6 GB.
Evaluation Measurements:

2.14 Android Adware and General Malware Dataset

Cited papers: [22]
Description: They installed the Android applications on the real device and captured its network traffic.
Research Goals:
Data Structure: The dataset is generated from 1900 applications with the following three categories: Adware (250 apps), General Malware (150 apps), and Benign (1500 apps). The dataset consists of the following: 1) .pcap files âĂŞ the network traffic of both the malware and benign (20% malware and 80% benign);2).csv files - the list of extracted network traffic features generated by the CICflowmeter.
Evaluation Measurements:

2.15 Android Malware Dataset (CICAndMal2017)

Cited papers: [38]
Description: They run both malware and benign applications on real smartphones to avoid runtime behavior modification of advanced malware samples that are able to detect the emulator environment.
Research Goals:
Data Structure: We collected more than 10,854 samples (4,354 malware and 6,500 benign) from several sources. We have collected over six thousand benign apps from Googleplay market published in 2015, 2016, 2017. We installed 5,000 of the collected samples (426 malware and 5,065 benign) on real devices. Our malware samples in the CICAndMal2017 dataset are classified into four categories: Adware, Ransomware, Scareware, SMS Malware. Our samples come from 42 unique malware families.
Evaluation Measurements:

2.16 CSE-CIC-IDS2018 on AWS

Cited papers:
Description: In CSE-CIC-IDS2018 dataset, we use the notion of profiles to generate datasets in a systematic manner, which will contain detailed descriptions of intrusions and abstract distribution models for applications, protocols, or lower level network entities. These profiles can be used by agents or human operators to generate events on the network. Due to the abstract nature of the generated profiles, we can apply them to a diverse range of network protocols with different topologies. Profiles can be used together to generate a dataset for specific needs.
Research Goals:
Data Structure: The dataset has been organized per day. For each day, we recorded the raw data including the network traffic (Pcaps) and event logs (windows and Ubuntu event Logs) per machine. In features extraction process from the raw data, we used the CICFlowMeter-V3 and extracted more than 80 traffic features and saved them as a CSV file per machine.
Evaluation Measurements:

2.17 CAIDA

Cited papers: [38] [5] [27]
Description: This organization has three different datasets, the CAIDA OC48, which includes different types of data observed on an OC48 link in San Jose, the CAIDA DDOS, which includes one-hour DDoS attack traffic split of 5-minute pcap files, and the CAIDA Internet traces 2016, which is passive traffic traces from CAIDAâĂŹs Equinix-Chicago monitor on the High-speed Internet backbone
Research Goals:
Data Structure: The CAIDA dataset contains 5 minutes (i.e., 300 s) of anonymized traffic obtained during a DDoS attack on August 4, 2007. These traffic traces store only attack traffic to the victim and response from the victim; non-attack traffic has been removed as much as possible. It is a high-rate attack if there are more than 10,000 packets per second over the network, with 1000 attack packets per second covering 60% of the attack traffic. As a result, this is low-rate attack traffic.
Evaluation Measurements:

2.18 LBNL dataset

Cited papers: [2]
Description: The dataset is an enterprise traffic dataset collected at the edge router of the Lawrence Berkeley National Lab (LBNL). Attack traffic in this dataset mostly comprises high-rate background traffic and low-rate outgoing scans. Traffic in this dataset comprises packet-level incoming, outgoing and internallyrouted traffic streams at the LBNL edge routers.
Research Goals:
Data Structure: This dataset contains both background traffic and attack traffic. Attack traffic was isolated by identifying scans in the aggregate traffic traces. Scans were identified by flagging those hosts which unsuccessfully probed more than 20 hosts, out of which 16 hosts were probed in ascending or descending order. Malicious traffic mostly comprises failed incoming TCP SYN requests.

Evaluation Measurements:

2.19 Kyoto

Cited papers: [39] [36] [9]
Description: Kyoto2006+ dataset is obtained from a honeypot networks of Kyoto University. In the honeypot networks, several types of honeypots are deployed over 5 different networks which are inside and outside of Kyoto University. There are some different OS, network printers and home information appliances (e.g. TV, Video Recorder). Kyoto2006+ dataset deploys a mail server in the same network to collect for normal traffic.
Research Goals:
Data Structure: From traffic data of the network, Kyoto 2006+ dataset extracts 14 conventional features and 10 additional features for each session. The former 14 features are extracted based on KDDCup 1999 dataset that is widely used for performance evaluation in intrusion detection system. The latter 10 features are extracted for more effective investigation. For example, signature-based IDS alerts, Antivirus alerts, source IP address and port number, time the session was started and so on.
Evaluation Measurements:

2.20 Twente

Cited papers: [40]
Description: This dataset includes three services such as OpenSSH, Apache web server and Proftp using auth/ident on port 113 and captured data from a honeypot network by Netflow. There is some simultaneous network traffic such as auth/ident, ICMP, and IRC traffic, which are not completely benign or malicious. There is some simultaneous network traffic such as auth/ident, ICMP, and IRC traffic, which are not completely benign or malicious. Moreover, this dataset contains some unknown and uncorrelated alerts traffic.
Research Goals:
Data Structure: The data collection resulted in a 24 GB dump file containing 155.2 million packets.
Evaluation Measurements:

2.21 UMASS

Cited papers: [28] [32]
Description: The dataset includes trace files, which are network packets, and some traces on wireless applications. It has been generated using a single TCP-based download request attack scenario.
Research Goals:
Data Structure:
Evaluation Measurements:

2.22 Drebin Dataset

Cited papers: [1] [42]
Description: In particular, we have acquired an initial dataset of 131,611 applications comprising benign as well as malicious software. The samples have been collected in the period from August 2010 to October 2012. In detail, the dataset contains 96,150 applications from the GooglePlay Store, 19,545 applications from different alternative Chinese Markets, 2,810 applications from alternative Russian Markets and 13,106 samples from other sources, such as Android websites, malware forums and security blogs. Additionally, the dataset includes all samples from the Android Malware Genome Project.
Research Goals:
Data Structure: The final dataset contains 123,453 benign applications and 5,560 malware samples.
Evaluation Measurements:

2.23 HTTP DATASET CSIC 2010

Cited papers: [42]
Description: The HTTP dataset CSIC 2010 contains the generated traffic targeted to an e- Commerce web application developed at our department. In this web application, users can buy items using a shopping cart and register by providing some personal information. As it is a web application in Spanish, the data set contains some Latin characters.
Research Goals:
Data Structure: The dataset is generated automatically and contains 36,000 normal requests and more than 25,000 anomalous requests. The HTTP requests are labeled as normal or anomalous and the dataset includes attacks such as SQL injection, buffer overflow, information gathering, files disclosure, CRLF injection, XSS, server side include, parameter tampering and so on.
Evaluation Measurements:

2.24 AWID

Cited papers: [20]
Description: This dataset is a publicly available collection of sets of data in easily distributed format, which contain real traces of both normal and intrusive 802.11 traffic.
Research Goals:
Data Structure: The AWID collection of datasets is comprised of two equal sets which defer merely on the labeling method (AWID-CLS, AWID-ATK). The first one is labeled according to the classification introduced in Section III-E (4 classes), while the latter follows a more detailed classification based on the actual attacks (16 classes).
Evaluation Measurements:

2.25 Coburg Intrusion Detection Data Sets(CIDDS)

Cited papers: [34] [33]
Description: They emulate a small business environment which includes several clients and typical servers. Network traffic is generated by scripts which emulate typical user activities like surfing the web, writing emails, or printing documents on the clients. These scripts follow some guidelines to ensure that the user behaviour is as realistic as possible, also with respect to working hours and

lunch breaks. The generated network traffic is recorded in unidirectional NetFlow format. For generating malicious traffic, attacks like Denial of Service, Brute Force, and Port Scans are executed within the network. Since origins, targets, and timestamps of executed attacks are known, labelling of recorded NetFlow data is easily possible. For inclusion of actual traffic, which has its origin outside the OpenStack environment, an external server with two services is deployed. This server has a public IP address and is exposed to real and up-to-date attacks from the internet.
Research Goals:
Data Structure: Nearly 32 million flows were captured from which around 31 million flows were captured in the OpenStack environment. Overall, we exploited 92 attacks within the four weeks.
Evaluation Measurements:

2.26 Comprehensive, Multi-Source Cyber-Security Events

Cited papers: [19] [31]
Description: This data set represents 58 consecutive days of de-identified event data collected from five sources within Los Alamos National LaboratoryâĂŹs corporate, internal computer network. The data sources include Windows-based authentication events from both individual computers and centralized Active Directory domain controller servers; process start and stop events from individual Windows computers; Domain Name Service (DNS) lookups as collected on internal DNS servers; network flow data as collected on at several key router locations; and a set of well-defined red teaming events that present bad behavior within the 58 days.
Research Goals:
Data Structure: In total, the data set is approximately 12 gigabytes compressed across the five data elements and presents 1,648,275,307 events in total for 12,425 users, 17,684 computers, and 62,974 processes.
Evaluation Measurements:

2.27 User-Computer Authentication Associations in Time

Cited papers: [15]
Description: This anonymized data set encompasses 9 continuous months and represents 708,304,516 successful authentication events from users to computers collected from the Los Alamos National Laboratory (LANL) enterprise network.
Research Goals:
Data Structure: Each authentication event is on a separate line in the form of "time,user,computer" and represents a successful authentication by a user to a computer at the given time. There are 11,362 users within the data set, and 22,284 computers. Timestamps, with a resolution of 1 second, start at an epoch 1 and all subsequent times are an offset from this epoch. The time frame of the actual data collection is not provided to enhance the anonymization of the data. The values are comma delimited. The data is available both as as one single file with 708,304,516 text lines or 9 files each with 30 days of events.
Evaluation Measurements:

2.28 Unified Host and Network Dataset

Cited papers: [44] [43]
Description: The Unified Host and Network Dataset is a subset of network and computer (host) events collected from the Los Alamos National Laboratory enterprise network over the course of approximately 90 days. The host event logs originated from most enterprise computers running the Microsoft Windows operating system on Los Alamos National Laboratory's (LANL) enterprise network. The network event data originated from many of the internal enterprise routers within the LANL enterprise network.
Research Goals:
Data Structure: The raw network flow data consisted of NetFlow V9 records that were exported from the core network routers to a centralized collection server. While V9 records can contain many different fields, only the following are considered: StartTime, EndTime, SrcIP, DstIP, Protocol, SrcPort, DstPort, Packets and Bytes. All windows host log data records will contain several attributes like EventID, LogHost, Time, LogonType, etc.
Evaluation Measurements:

2.29 Stratosphere CTU-13

Cited papers: [13]
Description: The CTU-13 is a dataset of botnet traffic that was captured in the CTU University, Czech Republic, in 2011. The goal of the dataset was to have a large capture of real botnet traffic mixed with normal traffic and background traffic. The CTU-13 dataset consists in thirteen captures (called scenarios) of different botnet samples. On each scenario we executed a specific malware, which used several protocols and performed different actions.
Research Goals:
Data Structure: After capturing the packets, the dataset was preprocessed and converted to a common format for the detection-methods. The format selected was the NetFlow file standard which is considered the ad-hoc standard for network data representation. The conversion from pcap files to NetFlow files was done in two steps using the Argus software suiteThese final NetFlow files were composed of the following fields: Start Time, End Time, Duration, Source IP address, Source Port, Direction, Destination IP address, Destination Port, State, SToS, Total Packets and Total Bytes.
Evaluation Measurements:

2.30 Detecting Malicious URLs

Cited papers: [23] [25] [24]
Description: For benign URLs, they used two data sources. One is the DMOZ Open Directory Project, which is a directory whose entries are vetted manually by editors. The second source of benign URLs was the random URL selector for YahooâĂŹs directory. They also drew from two sources for URLs to malicious sites: PhishTank and Spamscatter. PhishTank is a blacklist of phishing URLs consisting of manually-verified user contributions. Spamscatter is a spam collection infrastructure from which we extract URLs from

the bodies of those messages.
Research Goals:
Data Structure: The data set consists of about 2.4 million URLs (examples) and 3.2 million features.
Evaluation Measurements:

2.31 UNM Intrusion Detection Dataset

Cited papers: [46]
Description: Each trace is the list of system calls issued by an lpr process from the beginning of its execution to the end. There are 182 different system calls in the dataset.
Research Goals:
Data Structure: It consists of 4, 298 normal traces and 1,001 intrusion traces.
Evaluation Measurements:

2.32 Apache JIRA and Chromium Dataset

Cited papers: [30] [29]
Description: They use a total of five projects: four from Apache JIRA and a subset of bug reports from the Chromium project. Table The Chromium dataset comes from the 2011 mining challenge of the Mining Software Repositories conference.
Research Goals:
Data Structure: There are six kinds of high impact bugs reports in the datasets Apache JIRA. These include, surprise, dormant, blocking, security, performance and breakage. In Chromnium project, security bugs are labelled as Bug-Security when they are submitted to the system. Each row in Apache JIRA project represents a bug report and the columns are features of the reports such as bug id, title, description, and date and time a report was submitted and fixed. Chromium's html files are converted into a single CSV file with the column headers, id, date, report, and security.
Evaluation Measurements:

No.	Dataset	Year	Source	Accessible	Data Type	Percentage of abnormal data	Research Goals of Sample Cited Papers
1	ADFA-LD	2013	UNSW	Open	System Call Traces	12.13%	Used for evaluation of anomaly detection systems
2	CDX	2009	NSA, US Military Academy West Point	Open	TCP Dumps and Snort IDS Logs	-	Used for evaluating network intrusion models
3	KDD 98&99	1998-1999	University of California, Irvine	Open	Network Traffic	19.84%	Used for evaluation of anomaly detection methods
4	DARPA	1998-2000	DARPA & MIT Lincoln Labs	Open	Network Traffic	-	Used for the purpose of training and testing the intrusion detectors
5	ISCXIDS2012	2012	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	2.8%	Used for evaluation of intrusion detection systems
6	CICIDS2017	2017	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	-	Designed for evaluation of network attack detectors
7	NSL-KDD dataset	2009	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	48.12%	Designed to evaluate network intrusion detection systems
8	ISCXVPN2016	2016	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	-	Used to evaluate a flow-based classification method
9	Botnet dataset	2014	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	44.56%	Used to evaluate botnet detection systems
10	Android validation dataset	2014	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Android Apps	-	Used to test approach for the detection of Android app similarity
11	Android Botnet dataset	2015	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	URLs in Apk	10.95%	Used to evaluate the detection of botnets in Android
12	ISCXTor2016	2016	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	-	Used to evaluate the network traffic classifier
13	CIC DoS dataset	2017	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	-	Used to evaluate the detection approach for application layer DoS attacks
14	Android Adware and General Malware Dataset	2017	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	21.05%	Used to test an Android Malware Detection Model
15	CICAndMal2017	2017	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	7.76%	Used to evaluate anomaly-based intrusion detection approach
16	CSE-CIC-IDS2018 on AWS	2018	Canadian Institute for Cybersecurity University of New Brunswick	Request by email	Network Traffic	-	Used to evaluate network-based anomaly detectors
17	CAIDA	2002-2016	Center of Applied Internet Data Analysis	Open	Network Traffic	10%	Used to evaluate DDoS detector
18	LBNL	2004-2005	Lawrence Berkeley National Laboratory and ICSI	Open	Network Traffic Headers	28.33%	Used to evaluate the anomaly detectors
19	Kyoto 2006+ Dataset	2006-2009	Kyoto University	Open	Network Traffic	46.24%	Used for evaluation of NIDSs
20	Twente	2009	University of Twente	Open	Network Traffic	-	Used to evaluate a flow-based intrusion detection methods
21	UMASS	2011	University of Massachusetts	Open	Network Traffic	-	Used to evaluate network intrusion detection systems
22	Drebin	2010-2012	University of Gottingen	Request by email	Android Apps	4.3%	Used for detection of Android malware through static analysis
23	HTTP DATASET CSIC	2010	Spanish Research National Council	Open	Web requests	49.02%	Used for the testing of web attack protection systems
24	AWID	2016	University of the Aegean	Request by email	Wireless Traffic	3.32%	Used to test proposed classification algorithms
25	CIDDS	2017	Coburg University of Applied Sciences	Open	Network Traffic	5.29%	Used to evaluate anomaly-based intrusion detection systems
26	Comprehensive, Multi-Source Cyber-Security Events	2015	Los Alamos National Laboratory	Open	Event logs and authentication logs	-	Used to evaluate anomaly detection of events
27	User-Computer Authentication Associations in Time	2014	Los Alamos National Laboratory	Open	Authentication event logs	-	Used to model authentication data with a graph-based approach
28	Unified Host and Network Dataset	2017	Los Alamos National Laboratory	Open	Host and Network event logs	-	Used to evaluate anomaly detection systems
29	Stratosphere CTU-13	2011	Stratosphere Lab	Open	Network Traffic	2.34%	Used to compare three different botnet detection methods
30	Detecting Malicious URLs	2008	UCSD	Open	URLs	57.74%	Used to be applied in online learning algorithms to classify and predict malicious URLs
31	UNM Dataset	2004	University of New Mexico	Open	System Calls	19.14%	Used to evaluate intrusion detection systems
32	Apache JIRA and Chromium Dataset	2011	Mining Software Repository Conference	Open	Bug reports	0.5%	Used to test text-based prediction model for security bugs

Table 3: A list of existing security datasets

REFERENCES

[1] Daniel Arp, Michael Spreitzenbarth, Malte Hubner, Hugo Gascon, Konrad Rieck, and CERT Siemens. 2014. DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket.. In *Ndss*, Vol. 14. 23–26.

[2] Ayesha Binte Ashfaq, Maria Joseph Robert, Asma Mumtaz, Muhammad Qasim Ali, Ali Sajjad, and Syed Ali Khayam. 2008. A comparative evaluation of anomaly detectors under portscan attacks. In *International Workshop on Recent Advances in Intrusion Detection*. Springer, 351–371.

[3] Jean Paul Barddal, Heitor Murilo Gomes, and Fabrício Enembreck. 2015. SNCStream: A social network-based data stream clustering algorithm. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 935–940.

[4] Elaheh Biglar Beigi, Hossein Hadian Jazi, Natalia Stakhanova, and Ali A Ghorbani. 2014. Towards effective feature selection in machine learning-based botnet detection approaches. In *Communications and Network Security (CNS), 2014 IEEE Conference on*. IEEE, 247–255.

[5] Monowar H Bhuyan, DK Bhattacharyya, and Jugal K Kalita. 2015. An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection. *Pattern Recognition Letters* 51 (2015), 1–7.

[6] Monowar H Bhuyan, Dhruba K Bhattacharyya, and Jugal K Kalita. 2015. Towards Generating Real-life Datasets for Network Intrusion Detection. *IJ Network Security* 17, 6 (2015), 683–701.

[7] Carson Brown, Alex Cowperthwaite, Abdulrahman Hijazi, and Anil Somayaji. 2009. Analysis of the 1999 darpa/lincoln laboratory ids evaluation data with netadhict. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*. IEEE, 1–7.

[8] Song Chen and Vandana P Janeja. 2014. Human perspective to anomaly detection for cybersecurity. *Journal of Intelligent Information Systems* 42, 1 (2014), 133–153.

[9] Roshan Chitrakar and Chuanhe Huang. 2012. Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and naive Bayes classification. In *Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on*. IEEE, 1–5.

[10] Gideon Creech and Jiankun Hu. 2013. Generation of a new IDS test dataset: Time to retire the KDD collection. In *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*. IEEE, 4487–4492.

[11] Gideon Creech and Jiankun Hu. 2014. A semantic approach to host-based intrusion detection systems using contiguousand discontiguous system call patterns. *IEEE Trans. Comput.* 63, 4 (2014), 807–819.

[12] Gerard Draper-Gil, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. 2016. Characterization of Encrypted and VPN Traffic using Time-related. In *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*. 407–414.

[13] Sebastian Garcia, Martin Grill, Jan Stiborek, and Alejandro Zunino. 2014. An empirical comparison of botnet detection methods. *computers & security* 45 (2014), 100–123.

[14] Hugo Gonzalez, Natalia Stakhanova, and Ali A Ghorbani. 2014. Droidkin: Lightweight detection of android apps similarity. In *International Conference on Security and Privacy in Communication Systems*. Springer, 436–453.

[15] Aric Hagberg, Nathan Lemons, Alex Kent, and Joshua Neil. 2014. Connected components and credential hopping in authentication graphs. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*. IEEE, 416–423.

[16] I Homoliak, M Barabas, P Chmelar, M Drozd, and P Hanacek. 2013. ASNM: Advanced security network metrics for attack vector description. In *Proceedings of the International Conference on Security and Management (SAM)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1.

[17] Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A Ghorbani. 2017. Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. *Computer Networks* 121 (2017), 25–36.

[18] Andi Fitriah Abdul Kadir, Natalia Stakhanova, and Ali Akbar Ghorbani. 2015. Android botnets: What urls are telling us. In *International Conference on Network and System Security*. Springer, 78–91.

[19] Alexander D. Kent. 2015. Cybersecurity Data Sources for Dynamic Network Research. In *Dynamic Networks in Cybersecurity*. Imperial College Press.

[20] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Stefanos Gritzalis. 2016. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 184–208.

[21] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. 2017. Characterization of Tor Traffic using Time based Features.. In *ICISSP*. 253–262.

[22] Arash Habibi Lashkari, Andi Fitriah A Kadir, Hugo Gonzalez, Kenneth Fon Mbah, and Ali A Ghorbani. 2017. Towards a Network-Based Framework for Android Malware Detection and Characterization. In *Proceeding of the 15th international conference on privacy, security and trust*.

[23] Justin Ma, Alex Kulesza, Mark Dredze, Koby Crammer, Lawrence Saul, and Fernando Pereira. 2010. Exploiting feature covariance in high-dimensional online learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 493–500.

[24] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1245–1254.

[25] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Identifying suspicious URLs: an application of large-scale online learning. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 681–688.

[26] John McHugh. 2000. 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 (TISSEC)* 3, 4 (2000), 262–294.

[27] David Moore, Colleen Shannon, Douglas J Brown, Geoffrey M Voelker, and Stefan Savage. 2006. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)* 24, 2 (2006), 115–139.

[28] Joshua Ojo Nehinbe. 2011. A critical evaluation of datasets for investigating IDSs and IPSs researches. In *Cybernetic Intelligent Systems (CIS), 2011 IEEE 10th International Conference on*. IEEE, 92–97.

[29] Masao Ohira, Yutaro Kashiwa, Yosuke Yamatani, Hayato Yoshiyuki, Yoshiya Maeda, Nachai Limsettho, Keisuke Fujino, Hideaki Hata, Akinori Ihara, and Kenichi Matsumoto. 2015. A dataset of high impact bugs: Manually-classified issue reports. In *Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference on*. IEEE, 518–521.

[30] Fayola Peters, Thein Tun, Yijun Yu, and Bashar Nuseibeh. 2018. Text Filtering and Ranking for Security Bug Report Prediction. *IEEE Transactions on Software Engineering* (2018), Early–Access.

[31] Mir Mehedi A Pritom, Chuqin Li, Bill Chu, and Xi Niu. 2017. A Study on Log Analysis Approaches Using Sandia Dataset. In *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, 1–6.

[32] Swagatika Prusty, Brian Neil Levine, and Marc Liberatore. 2011. Forensic investigation of the OneSwarm anonymous filesharing system. In *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 201–214.

[33] M Ring, S Wunderlich, D Grüdl, D Landes, and A Hotho. 2017. Creation of Flow-Based Data Sets for Intrusion Detection. *Journal of Information Warfare* 16, 4 (2017), 41–54.

[34] Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho. 2017. Flow-based benchmark data sets for intrusion detection. In *Proceedings of the in Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS). 1em plus 0.5 em minus*. 361–369.

[35] Benjamin Sangster, TJ O'Connor, Thomas Cook, Robert Fanelli, Erik Dean, Christopher Morrell, and Gregory J Conti. 2009. Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets.. In *CSET*.

[36] Masaaki Sato, Hirofumi Yamaki, and Hiroki Takakura. 2012. Unknown attacks detection using feature extraction from anomaly-based ids alerts. In *Applications and the Internet (SAINT), 2012 IEEE/IPSJ 12th International Symposium on*. IEEE, 273–277.

[37] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization.. In *ICISSP*. 108–116.

[38] Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A Ghorbani. 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *computers & security* 31, 3 (2012), 357–374.

[39] Jungsuk Song, Hiroki Takakura, Yasuo Okabe, Masashi Eto, Daisuke Inoue, and Koji Nakao. 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*. ACM, 29–36.

[40] Anna Sperotto, Ramin Sadre, Frank Van Vliet, and Aiko Pras. 2009. A labeled data set for flow-based intrusion detection. In *International Workshop on IP Operations and Management*. Springer, 39–50.

[41] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*. IEEE, 1–6.

[42] Carmen Torrano-Gimenez, Alejandro Perez-Villegas, and Gonzalo Alvarez. 2009. A self-learning anomaly-based web application firewall. In *Computational Intelligence in Security for Information Systems*. Springer, 85–92.

[43] Melissa JM Turcotte, Alexander D Kent, and Curtis Hash. 2017. Unified host and network data set. *arXiv preprint arXiv:1708.07518* (2017).

[44] M. J. M. Turcotte, A. D. Kent, and C. Hash. 2017. Unified Host and Network Data Set. *ArXiv e-prints* (Aug. 2017). arXiv:1708.07518

[45] Miao Xie, Jiankun Hu, Xinghuo Yu, and Elizabeth Chang. 2014. Evaluating host-based anomaly detection systems: Application of the frequency-based algorithms to adfa-ld. In *International Conference on Network and System Security*. Springer, 542–549.

[46] JingTao Yao, Songlun Zhao, and Lisa Fan. 2006. An enhanced support vector machine model for intrusion detection. In *International Conference on Rough Sets and Knowledge Technology*. Springer, 538–543.