Requirements for Training and Evaluation Dataset of Network and Host Intrusion Detection System

Petteri Nevavuori and Tero Kokkonen

Institute of Information Technology,
JAMK University of Applied Sciences,
Jyväskylä, Finland
{petteri.nevavuori, tero.kokkonen}@jamk.fi

Abstract. In the cyber domain, situational awareness of the critical assets is extremely important. For achieving comprehensive situational awareness, accurate sensor information is required. An important branch of sensors are Intrusion Detection Systems (IDS), especially anomaly based intrusion detection systems applying artificial intelligence or machine learning for anomaly detection. This millennium has seen the transformation of industries due to the developments in data based modelling methods. The most crucial bottleneck for modelling the IDS is the absence of publicly available datasets compliant to modern equipment, system design standards and cyber threat landscape. The predominant dataset, the *KDD Cup 1999*, is still actively used in IDS modelling research despite the expressed criticism. Other, more recent datasets, tend to record data only either from the perimeters of the testbed environment's network traffic or from the effects that malware has on a single host machine. Our study focuses on forming a set of requirements for a holistic Network and Host Intrusion Detection System (NHIDS) dataset by reviewing existing and studied datasets within the field of IDS modelling. As a result, the requirements for state-of-the-art NHIDS dataset are presented to be utilised for research and development of NHIDS applying machine learning and artificial intelligence.

Keywords: Intrusion Detection, Dataset, Cyber Security, Machine Learning, Artificial Intelligence.

1 Introduction

During the last few years, we have seen some great developments in both domains of machine learning and cyber security. Companies and researchers have investigated various ways to join these two domains with a goal of reliably detecting malicious activity in complex networked systems. The development of Intrusion Detection Systems (IDS) is one of the state-of-the-art domains where machine learning and cyber security are combined. However, research and development of IDS requires generating realistic network traffic and intentionally mixing it with malicious attack traffic.

Generally, Intrusion Detection Systems are classified in two distinct ways. One way is to classify them according to the machine learning task they aim to fulfil. These tasks are *supervised signature detection* utilising labelled data and *unsupervised anomaly detection* using the normally functioning system as the baseline for comparison [7]. Another way to distinguish between the IDSs is by their operational context. Systems that focus on inter-host connection analysis are collectively called Network Intrusion Detection Systems (NIDS) while systems developed for intra-host activity analysis are called Host Intrusion Detection Systems (HIDS) [19].

Although the research and development within IDSs has been constant, the availability of required training and evaluation datasets is weak and commonly used misuse and anomaly detection datasets have fallen out-of-date. In some cases, the datasets date back to ten or even twenty years. These aged but still commonly used datasets fail to capture the complexity of modern systems and the evolved cyber threats imposed on them. Although several datasets are publicly available for research, they have been quite extensively criticised for lacking representativeness and generalisability to modern complex environments [32, 9, 3, 31]. As stated in [25], datasets are often overly anonymised, they do not include modern threats and intrusions, have weak statistical characteristics or are not released because of the security reasons.

As these datasets are effectively just snapshots from their relevant time period, using them is bound to negatively affect the usability of threat detection models trained with these datasets. Because publicly available datasets have limitations, researchers have at times utilised existing operational systems for producing evaluation datasets. Such data, however, requires anonymisation for protecting systems and people using them, which imposes another burden on the researchers willing to make their datasets publicly available. This has led researchers to either build small-scale laboratory-like environments for producing and collecting data, or resort to disputed existing public datasets. Small-scale laboratory-like environments do not tend to mimic real world IT environments with multiple connected systems and users. Some of the datasets are also produced in environments that are outdated with regard to both the target systems and the attacks performed on them. It is also usual for the datasets to focus only either on the network or host data, which disallows the holistic modelling of the state of a system.

1.1 Motivation and Structure

Relevant and effective implementation of anomaly-based IDS models requires realistic data. There exists a risk of developing irrelevant and out-of-date models right from the beginning when using outdated and misleading data.

In this paper, we will review the prominent intrusion detection datasets to identify the necessary requirements for a modern IDS modelling dataset. The requirements are defined in terms of overall dataset composition, collected data features and the systems used to produce the data by comparing existing datasets. We differentiate between requirements for network flow data and host-based

data, and review the data collection methods for each data type. We will then aggregate this information to identify the requirements for a modern NHIDS training and evaluation dataset. As will be shown in Section 2, there is a deficiency of datasets providing a snapshot of the environment in terms of both network traffic and host-based activities.

The paper is organised as follows. In Section 2 we first review the datasets selected for our review. We then compare the datasets cited and used in cyber security ID modelling research. We then define the requirements for an IDS modelling dataset capable of serving the needs of both network and host detection in Section 3. In Section 4 we discuss the environment in which the dataset could be generated. Lastly, we conclude our study in Section 5.

2 Dataset Review

As there already exist several cyber security method or dataset reviews providing information about distinct characteristics and features of existing public datasets, we have focused just on a concise representation of the existing datasets. The overview of datasets is divided into two subsections. In the first subsection we describe the prominently used datasets. The criterion that we used for defining a dataset as prominently used was that it should be either well-known or at least cited in several studies related to cyber security threat detection modelling. We give brief explanations about the selected datasets and refer to sources providing more in-depth dissection of their contents. However, restricting ourselves to review only datasets with a track record of utilisation means also that some of the most recently generated datasets are not included in the study. In the second subsection, we compare the features present in distinct datasets in order to see where the datasets agree and diverge across distinct features.

2.1 General Overview

First we review common datasets used with network and host IDS modelling. As the focus of our study is on building a set of dataset requirements for NHIDS modelling, we focus on datasets providing network traffic and host-based data. As these two are seldom combined, we have gathered information about datasets from both contexts.

KDD Cup 1999 [29, 33] has been widely used for nearly two decades [36, 19]. The dataset was created for an IDS development competition containing aggregated and processed network and host data. *KDD Cup 1999* dataset is based on the *DARPA 1998* dataset [5]. Chattopadhyay et al. [6] stated that the *KDD Cup 1999* dataset has been predominantly used in intrusion detection studies. For these reasons, we have only included the *KDD Cup 1999* dataset and omitted the *DARPA* datasets. Although several researches indicate that the dataset of *KDD Cup 1999* is not realistic, that dataset is still used as a benchmark for new methods and results of IDS study [4].

NSL-KDD [30] builds upon the *KDD Cup 1999* dataset by addressing the problems of redundant records and imbalance of malicious samples. Because the dataset is effectively only a filtered version of its predecessor, the dataset describes an at least ten-year old IT environment with expired attacks although constructed in 2009. While the use of the dataset is trumped by the use of the *KDD Cup 1999* dataset [6], it is still present in multiple studies either as the training or evaluation dataset. In their survey of machine and deep learning method research within the context of intrusion detection, Xin et al. [36] discuss multiple studies either utilising the *NSL-KDD* only, partially or in conjunction with other datasets.

The *Sperotto* [27] dataset contains labelled network flow records collected from a honeypot installed at a public network location. The authors refrained from performing explicit attacks on the monitored host systems to have the dataset further conform to real-world conditions. After the initial collection of data, the researchers labelled the dataset utilising a correlation process based on domain knowledge. The flow-based intrusion detection method survey by Umer et al. [32] mentions several studies that use the dataset for evaluation purposes.

ISOT Botnet [24, 35] is a labelled botnet network flow dataset merged from a collection of publicly available datasets during the time of the dataset's collection. The merging of the datasets was performed by replaying the network traffic from selected distinct datasets within a testbed environment. However, no details are given about the processes used to construct the malicious samples. While *ISOT* is a broader set of intrusion detection datasets, the *Botnet* dataset from 2011 has been used in several detection method studies [32, 2]. The dataset contains flow data for botnets and background traffic, however, it lacks normal data [12].

CTU-13 [12] is a flow-based botnet dataset like the *ISOT Botnet*. It consists of malicious data from an infected network and legitimate background data from a university's network. The dataset consists of multiple distinct scenarios with varying statistics regarding malicious and benign traffic instead of a single collective dataset. The dataset has been used in studies focusing on botnet detection from network flow data [17, 32].

ISCX IDS 2012 [25] dataset attempts to mimic real-world network events with traffic profiles constructed for distinct protocols. The dataset consists of network traffic flows collected from a simulated testbed environment. The profiles are essentially combinations of probability distributions fitted to features extracted from distinct protocols. The distributions are used for network flow data generation. This way the dataset should conform to real-world distributions of normal and background network traffic. Malicious traffic was generated by deliberate attacks performed on the testbed environment. Like *CTU-13*, the dataset contains several scenarios with distinct malicious activities. The detailed survey by Mishra et al. [19] reports dataset usage in several IDS studies.

ADFA-LD [10, 9, 8] and *ADFA-WD* [8] anomaly detection datasets are discretely host-based and built out of tokenized system call sequences. The datasets following in the footsteps of the 1998 *UNM* [13] system call dataset were con-

Table 1. Overview of prominently used IDS datasets.

Dataset	Year	Network data	Host data	Labelled	Acquisition[a]	Scenarios[b]	Malicious methods[c]	Avg. samples	Sample units
KDD Cup 1999	1999	Y	Y	Y	S	1	22	5,2 M	TCP Packets
NSL-KDD	2009	Y	Y	Y	S	1	22	1,2 M	TCP Packets
Sperotto	2009	Y	-	Y	R	1	?	14,2 M	Flows
ISOT Botnet	2010	Y	-	Y	M	1	3 (5)	1,7 M	Flows
CTU-13	2011[d]	Y	-	Y	S/R	13	7 (9)	80 M	Flows
ISCX IDS 2012	2012	Y	-	-	S	7	4	2,5 M	Flows
ADFA-LD	2013	-	Y	-	R	1	1 (6)	5,3 K	Syscalls
ADFA-WD	2014	-	Y	-	R	1	6 (12)	5,8 K	Syscalls
UNSW-NB15	2015	Y	-	Y	S	1	9	2,5 M	Flows

[a] S=Simulation, M=Merging of existing datasets, R=Real traffic/data.
[b] Split datasets (training, validation and testing) are considered a single dataset.
[c] Unbracketed numbers imply methods of conducting attacks (i.e. botnets) and bracketed numbers reported attack vectors.
[d] Associated report was first submitted in 2013, but dataset's website states 2011.

structed from Ubuntu and Windows host systems separately and they aim to facilitate the use of sequence-based models in host intrusion detection. System call tokenization enables learning the host system's benign baseline in terms of sequences of system calls performed by benign applications. Deviation from this baseline can be considered anomalous and potentially malicious. The datasets are featured in multiple host-based IDS studies referenced in relevant reviews [36, 1].

UNSW-NB15 [21, 20] is a synthetic network flow dataset constructed by sampling, filtering and aggregating flow information to form two sets of artificially generated network traffic captures. The testbed environment which the dataset was constructed consisted of three virtual servers attached to a traffic generator. While the design schematics of the testbed imply incorporation of normal client-side traffic, no mentions of background traffic generation are given. The dataset was built with the intention of replacing the predominant *KDD Cup 1999* dataset and it is reported being used in studies [19].

The general overview of the selected datasets occurring prominently in studies is presented in Table 1. The table contains information about the general context of the dataset, labelling status, acquisition method, scenario count, the numbers of malicious methods and possibly reported distinct attack vectors, sample counts and sample types. There is a notable dispersion across sample counts, scenarios and the number of malicious methods found in the datasets. Only two of the datasets, the *ADFAs*, are distinctly host-based while the rest are predominantly network traffic datasets.

Table 2. Public availability and formats of IDS datasets.

Dataset	Public	Pcap	Argus	Weblog	Text	CSV	DB	Netflow
KDD Cup 1999	Y	-	-	-	Y	Y[a]	-	-
NSL-KDD	Y	-	-	-	Y	Y	-	-
Sperotto	-	-	-	-	-	-	Y	-
ISOT Botnet	Y	Y	-	-	-	-	-	-
CTU-13	Y	Y	Y	Y	-	Y	-	Y
ISCX IDS 2012	-	Y[b]	-	-	-	-	-	-
ADFA-LD	Y	-	-	-	Y	-	-	-
ADFA-WD	Y	-	-	-	Y[c]	-	-	-
UNSW-NB15	Y	Y	Y	Y	-	Y	-	-

[a] Data is comma-separated albeit text.
[b] Format inferred from dataset sizes.
[c] Executables can be handled as text.

While there is research citing that various datasets exist, the search for the availability of the datasets produced varying and surprising results. Intuitively, public availability should have a great impact on how well a dataset is received by the community of researchers. The format of source data is also an important factor. The rawer the data, the greater the effort to reproduce the ready-to-use dataset.

The availabilities and data formats of the datasets are provided in Table 2. Public availability means that the datasets are readily available for anyone to obtain without limits and the rest of the columns indicate file formats in which the dataset is provided. The most common formats are the raw network traffic captures in the *pcap* format, the text file and the comma-separated value (CSV) file.

2.2 Network Data

In this subsection, we will compare the features of the datasets. We will focus only on network traffic datasets, as only the *KDD Cup 1999* dataset contains host-based data in conjunction with similar *ADFA* datasets. The host-based datasets are discussed briefly in Subsection 2.3. As with varying formats, there is also variance of features present across datasets. Directly comparing the features across datasets just by using the feature labels is misleading, as the naming conventions vary. The features have to be thus compared by their descriptions.

The descriptions of features for network traffic datasets are given distinctly in the following sources:

- *KDD Cup 1999* and *NSL-KDD* network traffic features are found in [5], as the latter is a subset of the former.
- *Sperotto* is described with details about its features in [26].
- *ISOT Botnet* is described in [24].

Table 3. Common network traffic dataset features.

Features	KDD Cup 1999	UNSW-NB15	ISOT Botnet	Sperotto	CTU-13	ISCX IDS 2012
Bytes, source to dest	Y	Y	Y	Y	Y	Y
Protocol, type	Y	Y	Y	Y	Y	Y
Bytes, dest to source	Y	Y	Y	Y	Y	-
Packets, dest	-	Y	Y	Y	Y	Y
Packets, source	-	Y	Y	Y	Y	Y
Timestamp, start	-	Y	Y	Y	Y	Y
Communication duration	Y	Y	-	Y	Y	-
Port, dest	-	Y	Y	Y	Y	-
Port, source	-	Y	Y	Y	Y	-
IP, dest	-	Y	Y	Y	Y	-
IP, source	-	Y	Y	Y	Y	-
Service accessed	Y	Y	-	-	-	Y
Protocol, state	-	Y	-	-	Y	Y

– *CTU-13* feature descriptions were derived from data headers.
– *ISCX IDS 2012* is described in [25].
– *UNSW-NB15* is documented and described in [34].

After collecting the features of distinct datasets, they were matched according to the provided descriptions. Some of the features were easy to match either directly with their labels or using descriptions. Others had to be looked into at a more general level e.g. how the number of packets or bytes relate to a flow or a connection. Some datasets report the counts separately for source and destination, while others only report a total count for communications. Because these describe essentially the same feature, they were handled effectively as similar. The features found in at least three of the datasets are presented in Table 3 with additional information about the dataset where the feature is present.

There were a total of 76 unique features across the datasets. Only 13 of these were common at least between any three datasets. The most commonly recorded features were the type of the communication protocol, the timestamp information and the counts for bytes and packets per communication. The source and destination address information was present four times, as was the communication duration. Some datasets incorporated records of accessed service and the state of the communication protocol. Even though some features were present just in a single dataset, multiple datasets had derived and aggregated features present complementing the features straightforwardly extractable from the raw traffic. While the complementary features were rather unique, their existence is common.

2.3 Host Data

Our selection of datasets has three either completely dedicated or network traffic complementing occurrences of host-based data. These are the *KDD Cup 1999*, the *ADFA-LD* and the *ADFA-WD*. The former of these contains the most features describing host activities performed by malicious actors. However, as discussed in [10, 9, 8], the operating system (OS), namely Solaris, from which the host activities were tracked was not selected with a focus on high deployment on markets. This poses a problem on generalisability even without discussing the agedness of the OS. Host data features are also highly processed and aggregated in terms of binary occurrences of certain indicators, which leaves a little room for more diverse modelling; detailed information of OS processes is altogether missing from the *KDD Cup 1999* dataset.

The *ADFA* datasets developed one and half decades after the *KDD Cup 1999* aim to address these two main shortcomings. The suffixes in the names point to Linux (*LD*) and Windows (*WD*) datasets. The host operating systems deployed for the corresponding datasets were Ubuntu 11.04 and Windows XP SP2, both of which were widespread in terms of installations and use commercially at the time. The approach to forming the datasets differs from the *KDD Cup 1999* and rather follows in the steps of the late 1990's *UNM* [13] dataset using raw system calls for detecting anomalous behaviour within a host system [36, 9]. This in turn allows for the sequential modelling of the host's benign baseline for detecting anomalies.

3 Requirements for a NHIDS Dataset

As a result of the study, the requirements for a complete, consistent and unambiguous NHIDS dataset are presented as follows.

The dataset shall include network traffic data. It is safe to conclude from the recent and past reviews of intrusion detection methods that network traffic data is the predominant choice for trying to detect malicious activities aimed at IT environments [19, 14, 6, 18].

The dataset shall include host activity data. However, strictly sticking to flow-based approaches tends to leave a great amount of information outside the reach of available anomaly or signature detection methods. This is because the network communications are nowadays commonly encrypted, leaving the payload of the traffic unobserved [32]. Inability to access and dissect the payloads of the traffic for detecting semantic attacks can be considered the main drawback of the flow only approach [28]. Thus, enriching the dataset with host-based data allows the use of network traffic data without limiting the detection possibilities of payload-induced manifestations in the target environment and its subsystems.

The dataset shall contain multiple scenarios. A static dataset is also always bound to be only a snapshot in time capturing a limited representation of the observed phenomenon. In data domains, where the rate of change is slow, the effect of limited observations is somewhat negligible. In the case of IT systems

in general, the rate of change is high and observable within the time frames of single years. The direction of change, however, is not only limited in the direction of time, but also easily observable across the various implementations of IT environments. This in turn creates a dilemma for intrusion detection - how to determine if an intrusion detection methodology is valid outside the dataset it was tested upon? Providing variability in the dataset itself is a way to tackle this. Like *CTU-13* and *ISCX IDS 2012* datasets, building scenarios with distinct activity profiles is a good starting point [25]. However, malicious activities are not the only form of variance introduced to the target environments. When enterprise networks are considered, it is normal that the topology is bound to change at least on the edges due to recruitments of new employees. Thus, introducing variation in the form of varying the target environment is another way to build up scenarios.

The data shall be representative of the real-world circumstances. Otherwise the dataset is insignificant. This requires paying attention to collecting data from background and normal network traffic [25] as well as mundane activities of host systems [9, 8] regardless of whether the activities or traffic are simulated or not.

The format of the data shall enforce usability. Considering the format in which the dataset should be provided, the data should be as readily usable as it can be from the viewpoint of the user of the data. For example, if two methods are developed independently of each other with the same raw data, the comparability of the results is left heavily dependent on the processes of aggregating and collecting the dataset. Thus, while some details are inevitably lost in the process, at least the network dataset should be in a commonly utilised row data format, for example CSV. However, mimicking the approach of Creech et al. [8], host data should consist of the appropriately tokenized sequences of system calls. While the text format is sufficient for host data, providing the data in CSV would help in removing a preprocessing step required to convert the system call tokens to lists or arrays. There should also be verbose descriptions of the features recorded within the dataset.

4 Producing NHIDS Dataset in Cyber Range

Producing an NHIDS dataset cannot be done in real production environments and a small-scale laboratory environment cannot produce a dataset with the required complexity. One possibility is to use a Cyber Range environment for dataset production.

Cyber Range functions as a research, development, training and exercise environment for the domain of cyber security. It is a closed and controlled environment providing the capability to mimic the required networks and systems for the purposes of research and development and supporting cyber security training and exercises. There can be replicated representations of the required organisation's network, systems, tools, and simulated Internet with background traffic from applications and users. In the Cyber Range environment, it is risk-free to use various attacks and intrusions with the required scenario. [22, 11, 15]

For example, JAMK University of Applied Sciences has implemented Cyber Range called Realistic Global Cyber Environment (RGCE). RGCE mimics global Internet services with botnet-based traffic generation and attached organisation environments [16, 15]. RGCE is used in scenario-based data generation for the development of anomaly based NIDS applying machine learning [23].

5 Conclusions

In this paper, we reviewed prominent intrusion detection datasets to identify the necessary requirements for producing a network and host data combining state-of-the-art intrusion detection system (NHIDS) training and evaluation dataset. We previewed the datasets in terms of overall composition of the datasets, the appearance of common features across them and the method their generation. A distinction was made between network and host datasets due to the differences in the nature of the data.

We found out that the coexistence of network and host data within a single dataset is uncommon. While flow-based network traffic datasets form the majority of datasets utilised in IDS modelling research, the encryption policies enforced nowadays render flow-based approaches unable to grapple attack vectors present in network traffic payloads. This is mitigable with host activity inclusion, which has data about the effects of payloads injected to the target environment.

While there are several network traffic features sharing common ground across multiple datasets, utilising these might not be enough. We found that, even though having unique formulations, having aggregated and derived features enriches the network flow dataset.

Utilisation of a state-of-the-art Cyber Range environment would be a prominent future research subject for creating a real-world circumstance matching IDS dataset according to found requirements. As small-scale testbed environments fail to mimic the complexity of real-world IT environments, the use of a multi-faceted, holistic Cyber Range could prove out to be a prominent platform for generating a state-of-the-art NHIDS training and evaluation dataset compliant with modern equipment and design standards.

Acknowledgment

This research is partially funded by the Regional Council of Central Finland/Council of Tampere Region and European Regional Development Fund as part of the *New Business Innovations from Data-analytics* project of JAMK University of Applied Sciences Institute of Information Technology.

References

1. Abubakar, A.I., Chiroma, H., Muaz, S.A., Ila, L.B.: A Review of the Advances in Cyber Security Benchmark Datasets for Evaluating Data-Driven Based Intrusion

Detection Systems. In: Procedia Computer Science. vol. 62, pp. 221–227. Elsevier (jan 2015). https://doi.org/10.1016/j.procs.2015.08.443
2. Alejandre, F.V., Cortés, N.C., Anaya, E.A.: Feature selection to detect botnets using machine learning algorithms. In: 2017 International Conference on Electronics, Communications and Computers, CONIELECOMP 2017. pp. 1–7. IEEE (2017). https://doi.org/10.1109/CONIELECOMP.2017.7891834
3. Aviv, A.J., Haeberlen, A.: Challenges in Experimenting with Botnet Detection Systems. In: Proceedings of the 4th Conference on Cyber Security Experimentation and Test. pp. 6–6. CSET'11, USENIX Association, Berkeley, CA, USA (2011), http://dl.acm.org/citation.cfm?id=2027999.2028005
4. Bodström, T., Hämäläinen, T.: State of the Art Literature Review on Network Anomaly Detection with Deep Learning. In: Galinina, O., Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) Internet of Things, Smart Spaces, and Next Generation Networks and Systems. pp. 64–76. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01168-0_7
5. Buczak, A., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials **PP**(99), 1 (2015). https://doi.org/10.1109/COMST.2015.2494502
6. Chattopadhyay, M., Sen, R., Gupta, S.: A Comprehensive Review and meta-analysis on Applications of Machine Learning Techniques in Intrusion Detection. Australasian Journal of Information Systems **22**(2018), 1–27 (may 2018). https://doi.org/10.3127/ajis.v22i0.1667
7. Chio, C., Freeman, D.: Machine Learning and Security. OŔeilly Media, Inc., CA (2018)
8. Creech, G.: Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks. Ph.D. thesis (2013), http://handle.unsw.edu.au/1959.4/53218
9. Creech, G., Hu, J.: Generation of a new IDS test dataset: Time to retire the KDD collection. In: IEEE Wireless Communications and Networking Conference, WCNC. pp. 4487–4492. IEEE (apr 2013). https://doi.org/10.1109/WCNC.2013.6555301
10. Creech, G., Hu, J.: A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguousand Discontiguous System Call Patterns. IEEE Transactions on Computers **63**(4), 807–819 (2014). https://doi.org/10.1109/TC.2013.13
11. Ferguson, B., Tall, A., Olsen, D.: National Cyber Range Overview. In: 2014 IEEE Military Communications Conference. pp. 123–128 (Oct 2014). https://doi.org/10.1109/MILCOM.2014.27
12. García, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. Computers and Security **45**, 100–123 (2014). https://doi.org/10.1016/j.cose.2014.05.011
13. Hofmeyr, S.A., Forrest, S., Somayaji, A.: Intrusion detection using sequences of system calls. Journal of Computer Security **6**(3), 151–180 (jul 1998). https://doi.org/10.3233/JCS-980109
14. Husak, M., Komarkova, J., Bou-Harb, E., Celeda, P.: Survey of Attack Projection, Prediction, and Forecasting in Cyber Security. IEEE Communications Surveys & Tutorials (c), 1–1 (2018). https://doi.org/10.1109/COMST.2018.2871866
15. JAMK University of Applied Sciences, Institute of Information Technology, JYVSECTEC: RGCE Cyber Range. http://www.jyvsectec.fi/rgce/, accessed: 23 November 2018

16. Kokkonen, T., Puuska, S.: Blue Team Communication and Reporting for Enhancing Situational Awareness from White Team Perspective in Cyber Security Exercises. In: Galinina, O., Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) Internet of Things, Smart Spaces, and Next Generation Networks and Systems. pp. 277–288. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01168-0_26
17. Mathur, L., Raheja, M., Ahlawat, P.: Botnet Detection via mining of network traffic flow. Procedia Computer Science **132**, 1668–1677 (2018). https://doi.org/10.1016/j.procs.2018.05.137
18. Mishra, P., Pilli, E.S., Varadharajan, V., Tupakula, U.: Intrusion detection techniques in cloud environment: A survey. Journal of Network and Computer Applications **77**(January 2017), 18–47 (2017). https://doi.org/10.1016/j.jnca.2016.10.015
19. Mishra, P., Varadharajan, V., Tupakula, U., Pilli, E.S.: A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection. IEEE Communications Surveys & Tutorials **PP**(c), 1–1 (2018). https://doi.org/10.1109/COMST.2018.2847722
20. Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS). pp. 1–6 (Nov 2015). https://doi.org/10.1109/MilCIS.2015.7348942
21. Moustafa, N., Slay, J.: The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Information Security Journal **25**(1-3), 18–31 (2016). https://doi.org/10.1080/19393555.2015.1125974
22. National Institute of Standards and Technology NIST: Cyber Ranges. https://www.nist.gov/sites/default/files/documents/2018/02/13/cyber_ranges.pdf, accessed: 23 November 2018
23. Puuska, S., Kokkonen, T., Alatalo, J., Heilimo, E.: Anomaly-based Network Intrusion Detection using Wavelets and Adversarial Autoencoders (2018), accepted in the International Conference on Information Technology and Communications Security, SECITC, 8-9 November 2018. Will be published in Lecture Notes in Computer Science by Springer
24. Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., Felix, J., Hakimian, P.: Detecting P2P botnets through network behavior analysis and machine learning. In: 2011 Ninth Annual International Conference on Privacy, Security and Trust. pp. 174–180 (July 2011). https://doi.org/10.1109/PST.2011.5971980
25. Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers and Security **31**(3), 357–374 (may 2012). https://doi.org/10.1016/j.cose.2011.12.012
26. SimpleWiki: Labeled Dataset for Intrusion Detection. https://www.simpleweb.org/wiki/index.php/Labeled_Dataset_for_Intrusion_Detection, accessed: 19 November 2018
27. Sperotto, A., Sadre, R., Van Vliet, F., Pras, A.: A Labeled Data Set for Flow-Based Intrusion Detection. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 5843, pp. 39–50. Springer, Berlin, Heidelberg (oct 2009). https://doi.org/10.1007/978-3-642-04968-2_4
28. Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C., Pras, A., Stiller, B.: An overview of IP flow-based intrusion detection. IEEE

Communications Surveys and Tutorials **12**(3), 343–356 (2010). https://doi.org/10.1109/SURV.2010.032210.00054
29. Stolfo, S.J., Fan, W., Lee, W., Prodromidis, A., Chan, P.K.: Cost-based modeling for fraud and intrusion detection: results from the JAM project. In: Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. vol. 2, pp. 130–144 (Jan 2000). https://doi.org/10.1109/DISCEX.2000.821515
30. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. pp. 1–6. IEEE (jul 2009). https://doi.org/10.1109/CISDA.2009.5356528
31. Tavallaee, M., Stakhanova, N., Ghorbani, A.A.: Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **40**(5), 516–524 (sep 2010). https://doi.org/10.1109/TSMCC.2010.2048428
32. Umer, M.F., Sher, M., Bi, Y.: Flow-based intrusion detection: Techniques and challenges. Computers and Security **70**, 238–254 (2017). https://doi.org/10.1016/j.cose.2017.05.009
33. University of California, Irvine: KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed: 23 November 2018
34. University of New South Wales: The UNSW-NB15 Dataset Description. https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/, accessed: 19 November 2018
35. University of Victoria, ISOT Research Lab: Datasets. https://www.uvic.ca/engineering/ece/isot/datasets/, accessed: 23 November 2018
36. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., Wang, C.: Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access **6**, 35365–35381 (2018). https://doi.org/10.1109/ACCESS.2018.2836950