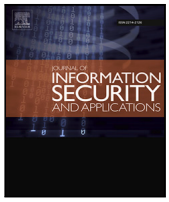
Journal of Information Security and Applications 80 (2024) 103689

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa



Benchmarking the benchmark — Comparing synthetic and real-world Network IDS datasets

Siamak Layeghy *, Marcus Gallagher, Marius Portmann

School of ITEE, University of Queensland, Brisbane, 4072, QLD, Australia

ARTICLE INFO

Keywords:
Network traffic characteristics
Feature distribution
Network Intrusion System (NIDS) dataset
Real-world NIDS dataset
Synthetic NIDS dataset
Machine learning benchmark dataset

ABSTRACT

Network Intrusion Detection Systems (NIDSs) are an increasingly important tool for the prevention and mitigation of cyber attacks. Over the past years, a lot of research efforts have aimed at leveraging the increasingly powerful models of Machine Learning (ML) for this purpose. A number of labelled synthetic datasets have been generated and made publicly available by researchers, and they have become the benchmarks via which new ML-based NIDS classifiers are being evaluated. Recently published results show excellent classification performance with these datasets, increasingly approaching 100 percent performance across key evaluation metrics such as Accuracy, F1 score, AUC, etc. Unfortunately, we have not yet seen these excellent academic research results translated into practical NIDS systems with such near-perfect performance. This motivated our research presented in this paper, where we analyse the statistical properties of the benign traffic in three of the more recent and relevant NIDS datasets, (CIC_IDS, UNSW_NB15, TON_IOT), by converting them into a common flow format. As a comparison, we consider two datasets obtained from real-world production networks, one from a university network and one from a medium size Internet Service Provider (ISP). Our results show that the two real-world datasets are quite similar among themselves in regards to most of the considered statistical features. Equally, the three synthetic datasets are also relatively similar within their group. However, and most importantly, our results show a distinct difference of most of the considered statistical features between the three synthetic datasets and the two real-world datasets. Since ML relies on the basic assumption of training and test datasets being sampled from the same distribution, this raises the question of how well the performance results of ML-classifiers trained on the considered synthetic datasets can translate and generalise to real-world networks. We believe this is an interesting and relevant question which provides motivation for further research in this space.

1. Introduction

Network Intrusion Detection Systems (NIDSs) are an important defence mechanism to protect computer networks against an increasing diversity, sophistication and volume of cyber attacks. Due to the tremendous progress in Machine Learning (ML) over the last few years, in particular Deep Neural Networks (DNNs), there has been a lot of recent research into leveraging the power of novel ML models for Network Intrusion Detection Systems. In particular supervised ML methods have shown great potential over traditional signature-based NIDSs. As with any supervised ML problem, the availability of high quality labelled datasets is absolutely critical. For many years, the KDD99 dataset [1] has been the most widely used benchmark dataset for the evaluation of NIDSs. However, it is well established that the dataset has significant limitations [2], the main one being its age, it was created over two decades ago. Given that Youtube, Facebook, Spotify, mainstream cloud computing and smartphones did not exist when the dataset was created, one can appreciate that the pattern of network traffic has undergone a profound change since then.

Furthermore, the type and sophistication of network attacks have undergone an equally dramatic change in the last 20+ years The need for more recent and relevant NIDS datasets has been clearly identified [2], and has lead to the development of a range of new datasets over the last few years. In contrast to ML application areas such as image classification, where high quality benchmark datasets can relatively easily be generated, this is a much harder problem in the context of NIDSs. Ideally, we would want datasets collected from real production networks, with realistic network patterns of benign traffic, together with a wide range of correctly labelled attack traffic. Since such ideal NIDS datasets are not readily available, researchers have recently developed a range of new synthetic datasets, which have become the new benchmarks. These synthetic datasets are typically

* Corresponding author.
E-mail address: siamak.layeghy@uq.net.au (S. Layeghy).

https://doi.org/10.1016/j.jisa.2023.103689

Available online 2 January 2024

generated in a controlled and relatively small simulation or testbed environment, where both benign traffic and attack traffic are created and labelled. Each of these datasets typically has its own dedicated feature set, which is collected and represented in a flow-based format. In the context of our paper, a flow is considered an aggregation of packets that share the same 5-tuple of source and destination IP address, source and destination transport layer port, and transport layer protocol.

Over the past few years, researchers have extensively used these synthetic datasets to evaluate a wide range of new proposed ML-based intrusion detection models and methodologies. Recently published results show increasingly excellent classification performance, approaching 100% across the key performance metrics, such as Accuracy, F1-score, etc. Consequently, one could assume that the problem of ML-based NIDS has been largely solved. Arguably, this is not quite the case, and the excellent results achieved in recently published academic research have not yet translated into practical, near-perfect intrusion detection systems deployed in real-world production networks. This apparent gap has motivated our research presented in this paper.

ML generally assumes that statistical properties of training data are the same as for testing data. Therefore, in order for the performance of an ML-classifier trained and evaluated on synthetic datasets to generalise and translate to a real network scenario, the statistical properties of both datasets would have to be similar. Our aim was therefore to compare the statistical properties of synthetic NIDS datasets with those of real world network traffic. In our analysis, we focused on benign (non-attack) traffic, due to the lack of attack labels in the production network traffic available to us. For our analysis, we have considered three recently published and widely used synthetic datasets UNSW_NB15 [2], CIC_IDS2017 [3] and TON_IOT [4]. The datasets contain 44 to 85 number of features in different formats. We further considered two datasets from large-scale real-world production networks, that we collected in 2019 and 2017. One dataset was obtained from a medium sized Australian ISP, and the other from the EAIT Faculty Network of the University of Queensland. The respective flow rates of the two networks were ~700 and ~100 flows per second. These real-world datasets were collected in Netflow/IPFIX format, which is widely available in real-world networks.

In order to enable the comparison of the five datasets, we required them to be in the same format. For this, we leveraged our previous work [5], where we converted synthetic NIDS datasets from their proprietary feature set and format to the standard Netflow/IPFIX format.[1] In [6], we argue the benefits of having a standard feature set and dataset format for NIDS. We also show that, somewhat surprisingly, ML-classifiers achieve higher performance using the Netflow feature set and format, compared to the original version.

For our comparison, we considered 6 practically relevant statistical features, such as flow duration, and packet size, plus a few IP address and layer 7 protocol related features. We compared the feature distributions via *box plots* and *Cumulative Distribution Function (CDF)* of features across the five considered datasets. We further quantified the distance of the different feature distributions using the Wasserstein metric [7]. Finally, in order to provide further intuitive insights in the different nature of the datasets, we calculated and visualised the embedding of four features into a 2 dimensional feature space, using four different embedding algorithms.

Our analysis provides the following key findings. The two real-world datasets, despite the fact that they had been collected from quite different networks, exhibit a high degree of similarity in the traffic feature distributions. Similarly, the synthetic datasets are quite similar amongst themselves in regards to most traffic features. However, and most interestingly, our analysis found a highly significant difference between the synthetic datasets and the real-world datasets in regards to most of the considered feature statistics.

To the best of our knowledge, this paper provides the first analysis of recent synthetic NIDS datasets and their comparison to real word traffic. We believe our results are relevant due to the extensive use of these synthetic datasets as a benchmark to evaluate ML-based NIDS models and algorithms, and they motivate future research into the development of new datasets that more closely match the properties of traffic in large scale real-world networks. This is an important goal in order to allow the translation and generalisation of the excellent NIDS performance achieved in academic research into NIDSs that are practically relevant and can be widely deployed in practical settings.

The rest of this paper is organised as follows. Section 2 discusses relevant related works. The three synthetic datasets along with the two real-world datasets are discussed in Section 3, and the feature selection/definition considerations along with the required prepossessing of datasets are explained in the next section. Section 5 provides comparisons between synthetic and real-world datasets, in terms of feature distribution box plots and CDFs, and Section 6 extends these comparisons by quantifying the differences of feature distributions using the Wasserstein distance metric. Section 7 illustrates the embeddings of the datasets using their dimensionality-reduced features. Section 8 discusses the limitations and strengths of this work and Section 9 concludes the paper.

2. Related works

Our discussion of related works consists of the following three parts, covered in the following corresponding subsections. First, we discuss works that investigate network traffic characteristics, i.e features of network traffic and their relationship to network behaviour. Secondly, we cover relevant works that aim to detect network traffic anomalies and discuss key network features in this context. Finally, we give an overview of the most closely related works to ours, i.e. papers that evaluate and analyse NIDS datasets.

2.1. Network traffic characteristics

A significant number of previous works, which discuss the network traffic characteristics, such as [8–10], mainly focus on traffic volume variations over time, to explain the network traffic characteristics. In [9] for instance, the authors use Principle Component Analysis (PCA) to investigate the origin–destination flows of a network as an essential part of network traffic modelling, in order to find solutions to a variety of problems including traffic engineering, capacity planning and anomaly detection. While time-series and time variations of network traffic volume play a significant role in network design and implementation, there are other network traffic features that are relevant when considering network traffic characteristics more generally. This has been illustrated in [11,12], where the authors investigate network traffic by analysing the statistical distributions of a broader range of traffic features such as flow duration, flow size, packet size and packet and flow inter-arrival time.

In [11], the authors studied the traffic characteristics of 10 distinct data centre networks with different organisational administrations such as university, enterprise, and cloud service providers. They studied the traffic patterns of these data centres by investigating the flow and packet-level properties of various layer-7 applications, and the impact of these applications on network congestion and link and network utilisation. For this, the authors investigated a range of packet and flow-based feature distributions, such as *number of flows per second*, *flow inter-arrival time*, *flow size*, *flow duration*, and *packet size*. In addition, unlike earlier works where the traffic volume variations over time was the main focus, this study also considered statistical feature distributions, including layer-7 features, for the purpose of network traffic characterisation.

[1] Datasets are available here https://staff.itee.uq.edu.au/marius/NIDS_datasets/

In [12] the authors investigated the traffic characteristics in data centre networks by collecting a petabyte of measurement data over a period of two months. They used three levels of network monitoring for their data collection, *SNMP counters*, sampled flow or sampled packet header level data, and deep packet inspection Since the paper's focus is on network traffic volume and network congestion identification, it mainly focuses on investigating network traffic patterns such as server, flow and bandwidth matrices. However, the paper also investigates features such as *flow duration*, *flow inter-arrival time* and the *number of destination servers per source server* to identify congestion patterns in their network.

2.2. Network features for anomaly detection

The next group of related works investigate network traffic characteristics in order to understand the normal network behaviour and detect anomalies. In this context, various statistics, measures and distributions of several network traffic features have been investigated.

In [13], the authors illustrated the importance of the distributions of flow and packet features, such as *IP addresses* and *ports*, in detecting and identifying a wide range of network anomalies. They show that clustering network traffic based on the distribution of these features can be utilised for anomaly detection. By investigating these feature distributions, the authors are able to detect different types of anomalies. The paper validates the proposed methods on data collected from two backbone networks (Abilene and Geant) and concludes that traffic feature distributions are a key ingredient in a general network anomaly detection framework.

The authors in [14] collected network traffic records of the same backbone networks as [13] (Abilene and Geant), to determine network features that affect the ability to detect anomalies. They recorded 3 weeks' of traffic and routing data from both networks, and detected three specific anomalies by applying a Kalman Filter on the entropy of the four considered network traffic features, i.e. *source and destination IP addresses* and *source and destination (L4) Ports*.

In [15], histograms of eight network traffic features have been investigated for the purpose of anomaly detection. These features include the *source and destination IP addresses*, *source and destination (L4) ports*, *TCP flags*, *(L4) protocol number*, *packet size*, and *flow duration*. The authors listed the possible benefits of each feature for detecting specific types of anomaly, and showed that a combination of features can reveal changes in the network traffic, which are otherwise not easily detectable. The evaluation was also based on real-world network traffic traces.

2.3. NIDS dataset evaluation

While there are many publicly available NIDS benchmark datasets, which have been widely used for the evaluation of new and existing machine learning based network intrusion detection algorithms, studies that evaluate these benchmark datasets themselves are relatively rare. This subsection discusses the key related works in this space.

The earliest studies to investigate NIDS datasets that we have found are [16,17]. Their main focus is the analysis and criticism of the DARPA dataset [18]. A key point mentioned in these works is in regards to the background traffic in the DARPA dataset, and the lack of statistical information about a comparison with real-world traffic. While the DARPA dataset itself and these two works are quite old, the main point made in the paper, i.e. the effect of background traffic on the detection performance, is still highly relevant, and is critical for the findings presented in our paper.

The next paper [19] tries to address the issues raised in the two previous studies relating the evaluation of the DARPA NIDS dataset. The paper uses two signature-based NIDSs, Snort [20] and Cisco IDS, along with two anomaly detection methods, in order to evaluate the DARPA dataset, based on the methodology proposed by MIT's Lincoln Laboratory [21]. The authors concluded that since the NIDSs available at the time were not capable of detecting attacks beyond what was included in the DARPA dataset, DARPA was still useful for NIDS evaluation.

There are a few studies such as [22,23] that investigate the KDD99 dataset [1], discuss its drawbacks and propose new datasets. In [22], the authors performed a statistical analysis of the KDD99 dataset and concluded that it has two major shortcomings that severely affect its suitability for the evaluation of NIDS algorithms including repeated/redundant records and its level of difficulty (very simple ML models can efficiently detect anomalies). The authors propose a new NIDS dataset, *NSL-KDD*, by selecting a subset of records from the KDD99 dataset that does not suffer from the mentioned shortcomings.

In [23], the authors identify three major issues related to the KDD99 dataset, lack of up-to-date network attacks, lack of up-to-date background traffic, and the distribution differences between the test and training sets. They propose a new dataset (UNSW_NB15) to address these shortcomings. However, the main method used to evaluate the KDD99 dataset is to compare the performance of different classifiers on the two datasets, which does not necessarily mean the new dataset has addressed shortcomings of KDD99 dataset.

The only study we found which provides a comprehensive evaluation on multiple NIDS datasets is [24]. The paper systematically evaluates 11 publicly accessible NIDS datasets, published between 1998 and 2016, and conclude that most of these datasets are out of date and unreliable for the evaluation of the NIDS algorithms. In order to identify the shortcomings of the existing datasets, the authors provide a framework consisting of 11 qualitative properties of a benchmark NIDS dataset. By assigning a point value for each property, they assign a total score for each considered dataset. These considered properties include complete network configuration, complete traffic, labelled dataset, complete interaction, complete capture, available protocols, attack diversity, anonymity, heterogeneity, feature set, and metadata. While this is the most comprehensive study of NIDS datasets, the evaluation of the benchmark datasets, and the framework, both are provided based on a high level qualitative approach, which is in contrast to the quantitative approach taken in our paper.

There are more recent works, such as [25,26], that compare the performance of different classifiers on a number of benchmark NIDS datasets. The papers also provide a high level comparison of the datasets, considering aspects such as percentage of protocol types, ratio of attack and benign traffic, etc. Another recent study of NIDS datasets [27] surveys a large number of NIDS datasets, focusing on high level information, without providing a statistical analysis of the features.

In summary, none of these works offer a quantitative and statistical analysis of synthetic NIDS benchmark datasets compared to real-world datasets. We believe our paper addresses this limitation and bridges the gap between the academic performance of ML-based NIDSs and their practical deployment in real-world networks.

3. Datasets explored in this study

Table 1 shows a list of the datasets investigated in this paper, consisting of three synthetic/ testbed-based NIDS benchmark datasets in their original format, their converted versions in NetFlow format, and two real-world datasets also in NetFlow format. The table shows the dataset type, size (number of flows), the percentage of attack flows, the number of features, the year the dataset was collected/published, and the tools used to generate/collect the dataset. The next three subsections explain these three groups of datasets.

3.1. Synthetic datasets (original format)

The synthetic/testbed-based datasets selected for this study are highly relevant and widely used NIDS benchmark datasets, which were

Table 1
Summary information of datasets studied in this paper.

Dataset type	Dataset	Attack flow ratio	Number of features	Year	Format	Generation tool
Synthetic (Original Format)	UNSW_NB15 [2]	12.65%	49	2015	proprietary	Argus/Bro
	CIC_IDS2017 [3]	28.16%	85	2017	proprietary	CIC FlowMeter
	TON_IOT [4]	96.44%	44	2019	proprietary	Security Onion/Bro
Synthetic (**NetFlow**)	Converted UNSW_NB15	**0**	20	2015	NetFlow	conversion by nProbe
	Converted CIC_IDS2017	**0**	20	2017	NetFlow	conversion by nProbe
	Converted TON_IOT	**0**	20	2019	NetFlow	conversion by nProbe
Real-World (**NetFlow**)	ISP [Collected by us]	**0**	20	2017	NetFlow	nProbe
	UQ [Collected by us]	**0**	20	2019	NetFlow	nProbe

published from 2015 till 2019. The first two datasets are among the most highly cited NIDS datasets, and the third dataset is a moderately new dataset. Further details on these datasets are provided in the following.

- UNSW_NB15 Dataset: This is the oldest among the three considered datasets, published in 2015 by the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) [2]. The paper introducing this dataset currently has 1300 citations, which highlights its relevance as an NIDS benchmark dataset. The dataset consists 2,540,044 flows, made up of 87.35% benign and 12.65% attack flows. Each flow record consists of 49 features generated by the Argus and Bro-IDS tools. The traffic was generated on testbed network, consisting of 9 different attack types, combined with normal (benign) background traffic. In addition to the Argus/Bro flow format, the dataset is also available in the form of raw *packet capture (pcap)* files.
- CIC_IDS2017 Dataset: This dataset has been generated by the Canadian Institute for Cybersecurity in 2017 [3]. The paper introducing this dataset is currently cited by 1021 other works. The main priority in generating this dataset is stated by authors as *having realistic background traffic*. They have used their own developed tool for network analysis, CICFlowMeter, to generate a dataset in which flows are labelled based on time stamp, source, and destination IPs, source and destination ports, protocols, plus a range of other extracted features, with a total of 85 features. The attack flows constitute %28.16 of total dataset and include 21 different network attacks, including DoS, DDoS, different scans, Botnet, and various types of Web and Infiltration attacks. The authors ran individual classes of attacks separately and provided the corresponding flow records in separate CSV files. Similarly, the benign traffic is provided in a separate file. This dataset is also published in the raw pcap format.
- TON_IOT Dataset: This dataset, published by ACCS in 2019 and referenced in [4], offers a more comprehensive range than their previous dataset (UNSW NB15 Dataset) [2]. It includes traffic records from their testbed network, IoT devices, and operating system logs. It includes 22,339,021 network flows expressed in 44 features extracted by Bro-IDS. The background/benign traffic constitutes 3.56% of flows in the dataset, and the remaining flows (96.44%) are made up of 9 different attack types. As for the other two synthetic datasets, this dataset is also published in the raw pcap format.

3.2. Synthetic datasets (NetFlow)

A statistical comparison in terms of their feature distributions, of the above datasets, in their original flow formats, is not possible due to the significant difference in feature sets. We have made the case for the use of a standard NIDS feature set in one of our previous works [6], and have discussed that this also allows the comparison of different ML-models across different NIDS benchmark datasets.

In order to enable a comparison of the different NIDS benchmark dataset, we convert them to a common format and feature set. This is possible, since the three considered synthetic datasets are also available in a raw pcap format, as already mentioned above.

We chose NetFlow as the common flow format, for a number of reasons. Firstly, it is the predominant industry standard for flow data collection in production networks, and tools for exporting, collecting and analysing data are readily available. Furthermore, as we have shown in [6], NIDS datasets represented in NetFlow format outperform the corresponding datasets in their original flow/feature representation in terms of ML-based network attack classification.

Next, we will discuss the NetFlow fields that we considered as the relevant flow features, for the purpose of comparison of datasets in this paper. Then we will briefly outline the dataset conversion procedure we used to convert the NIDS datasets from their original format and feature sets, to the common NetFlow-based feature set.

3.2.1. NetFlow features

Table 2 shows the list of NetFlow fields/features used for the statistical comparison of datasets in this paper. Despite NetFlow supporting a vast array of fields, it is important to note that many of these fields are protocol-specific, specifically tailored to certain protocols like S.I.P, lacking information about other protocols. we focused on a small subset of fields that we consider critical in the context of NIDS, based on previous studies of NIDS, and/or their role in network traffic volume and characteristics, as will be discussed in Section 5. We used NetFlow Version 9 (which is also the basis for IPFIX) for data collection/conversion, which allowed us to extract much broader information relating flows, such as the L7 PROTO field, indicating the layer 7 protocol of traffic.

3.2.2. Dataset format conversion

The conversion of multiple NIDS benchmark dataset to a common format and feature set is a critical requirement in order to perform a valid statistical comparison, in particular in regards to feature distributions. To the best of our knowledge, this has not been done yet in the context of NIDS benchmark datasets. The conversion of the datasets is enabled due to the availability of the considered labelled synthetic NIDS in the raw packet capture format (pcap), in addition to their original (proprietary) flow format. Our format conversion only considers benign traffic. This is due to the unavailability of labelled NIDS datasets from real-world production networks, which limits our statistical comparison to normal (benign) traffic. While we focus on the

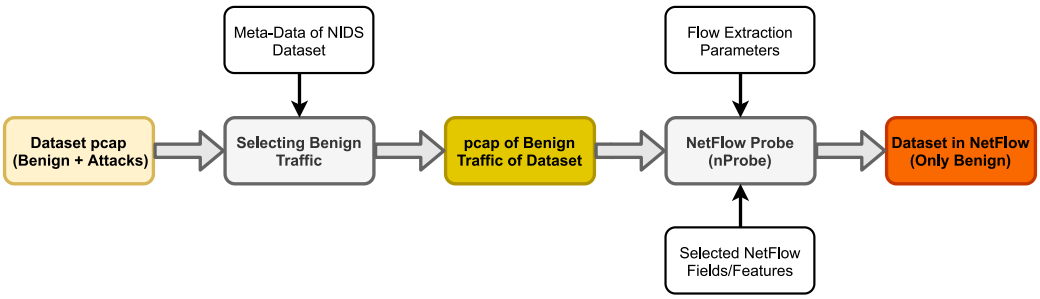


Fig. 1. The process of converting proprietary datasets into NetFlow format.

Table 2
List of the main NetFlow features used in this study for data collection and conversion.

NetFlow field	Definition
IPV4_SRC_ADDR	IPv4 source address
IPV4_DST_ADDR	IPv4 destination address
IN_PKTS	Number of incoming flow packets (src → dst)
OUT_PKTS	Number of Outgoing flow packets (dst → src)
IN_BYTES	Number of incoming flow bytes (src → dst)
OUT_BYTES	Number of outgoing flow bytes (dst → src)
FIRST_SWITCHED	System Up time (msec) of the first flow packet
LAST_SWITCHED	System Up time (msec) of the last flow packet
PROTOCOL	IP protocol byte
L7_PROTO	Layer 7 protocol (numeric)

format conversion of benign traffic, the conversion of the full datasets, including attack traffic, has been made publicly available in the context of our previous work [5].

The main tool utilised for the NIDS dataset format conversion is *nProbe* [28], an efficient tool used in real-world production networks for NetFlow exporting and collection. In addition to NetFlow exporting and collection, nProbe also makes it possible to read packet capture (pcap) files in streaming mode, simulating the real-time export of NetFlow records. This allows us generate NetFlow records from the pcap file.

Fig. 1 illustrates the overall process of converting these datasets from their proprietary formats into NetFlow. Before processing the pcap files via nProbe, the benign (attack free) parts of flows are selected, based on the metadata (labels) by the dataset publishers. In order to export NetFlow records, nProbe needs two sets of parameters. The first set of parameters, such as maximum time to export a flow, maximum idle time and buffer size, define when and on what conditions flow records are exported. The second set of nProbe parameters is the *export template*, which defines the NetFlow fields to be included in the generated NetFlow records. For the both sets of parameters, we used the same values/fields that had been used to extract our two real-world datasets. Details of all nProbe parameters used in this process are mentioned in [5].

3.3. Real-world datasets

In order to compare the statistical properties of the synthetic NIDS benchmark datasets, we collected Netflow records from two real-world production networks. The two datasets were collected from different organisations, with a very different customer base and hence varying types of network traffic. One dataset was collected from the network of the EAIT (Engineering, Architecture and Information Technology) faculty of the University of Queensland, and the other from a medium size Internet Service Provider (ISP) headquartered in Brisbane, Australia. These two datasets were recorded in two different time periods, 2017 and 2019, which overall provides us with a good degree of variability, and the confidence that the observed feature distributions cannot be attributed to a specific network setup.

- *Internet Service Provider (ISP):* The first real-world NetFlow dataset was obtained from an Australian Internet Service Provider (ISP) in the context of a research collaboration. The ISP has around 300 business customers and provides enterprise-grade managed services such as Internet, MPLS, VoIP, and cloud with dual point presences (PoPs) in all major Australian capital cities, as well as New Zealand, Hong Kong and the Philippines. We collected the NetFlow records of the entire backbone traffic via port mirroring on one of the routers, and running nProbe as a NetFlow exporter/collector on a server in the ISP's data centre. The collected dataset represents the entire backbone traffic (no sampling) for a duration of 30 days in June 2017, with a total of 162 GB of flow records.
- *University of Queensland (UQ):* The second real-world dataset NetFlow dataset was collected from the network of the faculty of Engineering, Architecture and Information Technology (EAIT) at the University of Queensland (UQ). The EAIT faculty consists of five schools including Architecture, Chemical Engineering, Civil Engineering, Information Technology & Electrical Engineering, and Mechanical & Mining Engineering. The nProbe software was used for NetFlow collection and export, with the identical flow extraction parameters and NetFlow features (export template) as in the case of the ISP dataset. The data collection at this site started in February 2019 and lasted for 50 days. The recorded data includes all the traffic flows in the monitored part of the network including all wired and wireless communication totally about 395 GB.

4. Feature considerations

4.1. Selecting attack-free parts of datasets

Network attacks and anomalies typically change the statistical distribution of many traffic features [13–15]. Accordingly, statistical distributions of features of the synthetic datasets that contain anomalies/attacks are expected to be different from our real-world datasets with unknown attack inclusion status. Consequently, since our aim is to compare the statistical characteristics of synthetic and real-world NIDS datasets, we need to make sure we select only attack-free records from the both groups of datasets.

This was easy in the case of the synthetic datasets since the metadata (label) for each flow is provided by dataset publishers. However, in the case of the real-world datasets this was a much harder task, which required meta-data collection and investigation along with manual inspection of the flow records.

Due to the time consuming nature of this task, it is virtually impossible to run this operation for the entire real-world datasets. Consequently, we instead selected a single representative day from each real-world dataset for our analysis.

Since Intrusion Detection and Prevention Systems (IDS/IPS), firewalls and advanced security and network monitoring appliances are utilised in both organisations to secure their networks, not many known

Table 3
List of Traffic features investigated in this study, along with their definition/formula.

Item	Feature	Definition/Formula
1	Flow duration	LAST_SWITCHED - FIRST-SWITCHED
2	Flow size	IN_BYTES + OUT_BYTES
3	Packet time	flow duration/(IN_PKTS + OUT_PKTS)
4	Packet size	flow size/(IN_PKTS + OUT_PKTS)
5	Number of destination Ips per source IP	$\lvert D_{x_i} \rvert$, with D_{x_i} as defined in Eq. (1)
6	Number of L7 protocols per source IP	$\lvert L_{x_i} \rvert$, with L_{x_i} as defined in Eq. (2)

attacks were expected in the parts of these networks that were used for the data collection. Nonetheless, we asked for the attack/anomaly logs from both the ISP and our university IT department, related to the entire periods of data collection. After investigating these records, a single day of flow records, with the least probability of attacks/anomalies was selected. In addition, a further rigorous manual inspection was performed to make sure all the selected records in the real-world datasets are benign (attack free), as much as possible. To inspect the flow records of the chosen representative day manually, we initially analysed the distribution of the features mentioned in Section 4.2. As will be shortly discussed, these features are related to the presence of attacks, and some of them have been used in many previous studies of network intrusion detection. Subsequently, flow records exhibiting a notable deviation from the distribution mean were further examined for evidence of attacks. This examination was carried out by a network security expert in conjunction with the existing network and server logs. During this thorough examination, the security expert meticulously reviewed network traffic patterns, scrutinised firewall logs, and cross-referenced any suspicious activities with known attack signatures to identify potential threats or anomalies. Furthermore, previous and subsequent flow records from the same source were analysed to identify any abnormal behaviour.

4.2. Feature selection

Table 3 lists the network traffic features that constitute the basis of analysis and comparison in this paper. The first column is the item number, the second column represents the feature name, and the last column shows the feature definition in terms of NetFlow fields. In the selection of these features, we considered their relevance in the detection of network anomalies and intrusions.

The first 4 features in Table 3, which have a value per flow records, directly or indirectly, have been used in most previous studies of network intrusion detection such as [8,11,29]. They are also the only 4 features that are common to all three synthetic NIDS datasets considered in this paper, which highlights their relevance. These 4 features can be briefly defined as

- *flow duration:* difference of the time when the last packet (LAST_SWITCHED) and first packet (FIRST_SWITCHED) of the flow have arrived into the switch.
- *flow size:* sum of the all flow bytes in the inward (IN_BYTES) and outward (OUT_BYTES) directions.
- *packet time:* average packet time computed by dividing flow duration by the whole number of flow packets in inward and outward directions.
- *packet size:* average packet size (in bytes) computed by dividing the flow size by the whole number of flow packets in inward and outward directions.

We have further added two more features that we consider relevant in this context. They do not have a value per flow and are calculated for the flow aggregations. Feature number 5 (*number of destination IP addresses per a source IP*) represents the number of distinct destination IP addresses that correspond to a single source IP address. This feature represents a critical aspect of the traffic matrix, i.e. the *fanout*. The last considered feature (*number of L7 protocols per source IP*), represents the number of distinct Layer 7 protocols that are associated with a specific source IP address. Again, this is a practically relevant feature related to the traffic matrix, and it provides an indication of the range of applications accessed by a single IP address.

For the more formal definition of features 5 and 6, we introduce some basic notation. We define a dataset $\mathcal{X} = \{x\}$ of NetFlow records, where x represents a single record, with N source IP addresses, M destination IP addresses, and K layer-7 protocol values. $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ and $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ represent the set of source IP addresses, destination IP addresses and L7-protocol values respectively. The set of all destination IP addresses of flows with the same source IP address, $\mathcal{D}_{\mathcal{X}_i}$, can be defined as

$$\begin{cases} \mathcal{D}_{\mathcal{X}_i} = \{d(x) |\ x \in \mathcal{X}_i\} \\ \\ \mathcal{X}_i = \{x |\ x \in \mathcal{X},\ s(x) = s_i\} \end{cases} \quad for\ i \in \{1, 2, \ldots, N\} \tag{1}$$

where $s(x)$ and $d(x)$ represent the source and destination IP address of flow x respectively. Accordingly, the number of destination IPs per source IP (feature number 5 in Table 3) can be defined as $|\mathcal{D}_{\mathcal{X}_i}|$.

Similarly, the set of all layer-7 protocols values of flows with the same source IP address, $\mathcal{L}_{\mathcal{X}_i}$, can be stated as

$$\begin{cases} \mathcal{L}_{\mathcal{X}_i} = \{l(x) |\ x \in \mathcal{X}_i\} \\ \\ \mathcal{X}_i = \{x |\ x \in \mathcal{X},\ s(x) = s_i\} \end{cases} \quad for\ i \in \{1, 2, \ldots, N\} \tag{2}$$

where $l(x)$ represents the layer-7 protocol of flow x. Hence, the number of L7 protocols per source IP (feature 6 in Table 3) is defined as $|\mathcal{L}_{\mathcal{X}_i}|$.

4.3. Intra-dataset feature variability

As discussed in Section 4.1, our real-world datasets include a single day of flow records, selected from the entire dataset of 30 or more days. This is done in order to keep the manual inspection of the flows manageable. As mentioned above, this time consuming inspection is necessary in order to provide a high degree of confidence that the dataset is indeed attack free.

Choosing a single day's data might raise the question of how representative the chosen day is in terms of the distributions of our considered flow features (Table 3) To address this, the intra-dataset feature variability, i.e. feature distribution within different full-day windows of the same real-world dataset, is investigated in this section. Fig. 2 shows the distribution of two of the considered features for individual days for both real-world datasets, with the data for the selected representative day (day 25 for UQ dataset, and day 33 for ISP dataset) shown in blue. In particular, Fig. 2-a and b show the distribution of *flow duration* and Fig. 2-c and d show the distribution of *flow size* for different days of the ISP and UQ datasets (sampled at 20 samples per minute). Similarly, Figures 2-e and f display packet time distribution, while Figures 2-g and h depict packet size distribution for various days in the ISP and UQ datasets.

The red circles, red crosses and grey lines/bars indicate the median, mean, and standard deviation (STD) of the specified features respectively. The y-axis is shown in logarithmic scale for the flow size feature (2-c and d) due to the presence of very large flows and hence the large range of flow size values. As can be seen in the figure, while there

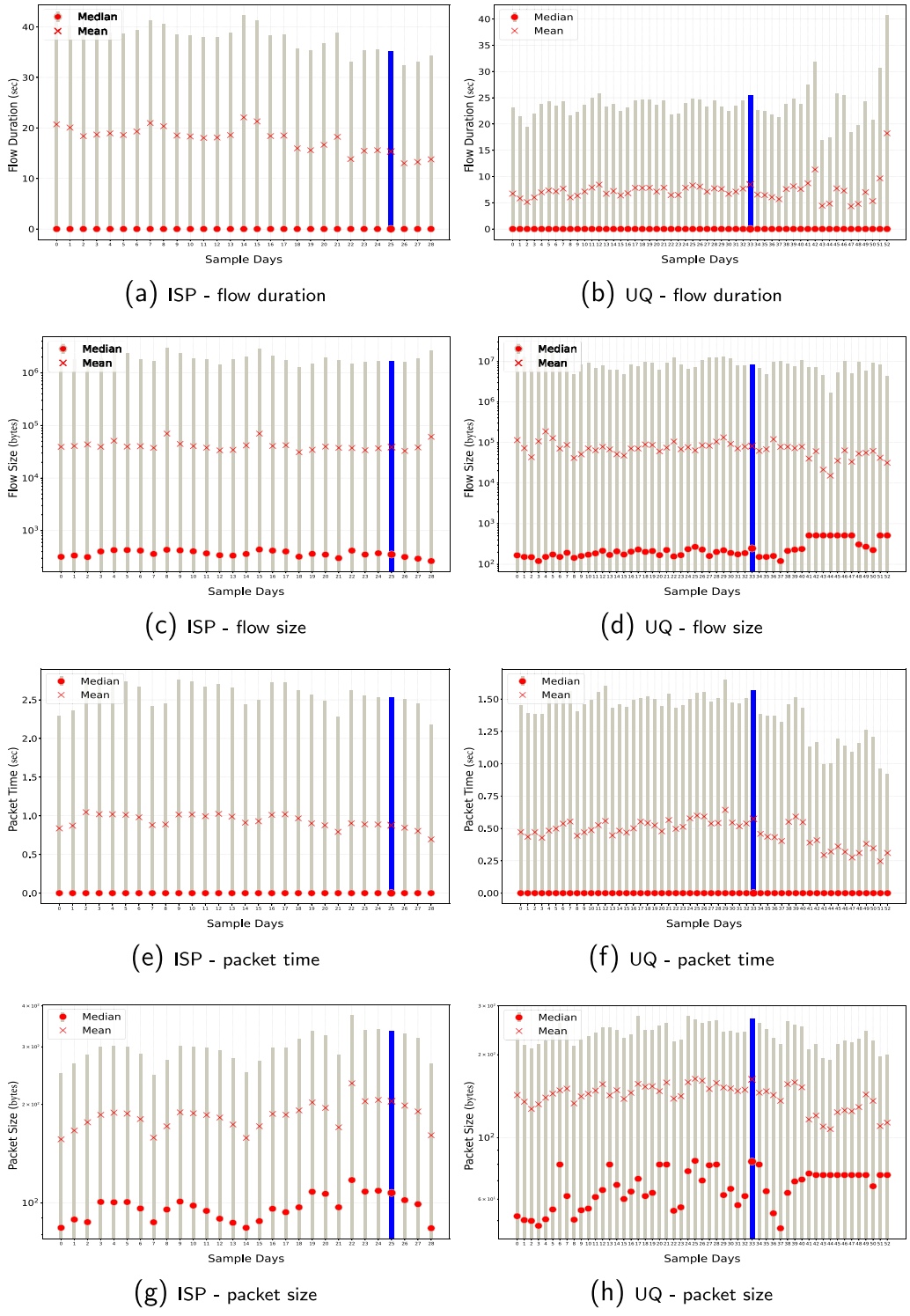


Fig. 2. Comparing the flow durations and flow sizes within the two real world datasets. (a) and (b) show the flow durations of ISP, and sampled UQ dataset, and (c) and (d) show their flow sizes correspondingly. (e) and (f) show the packet time for ISP and sampled UQ dataset, and (g) and (h) show the packet size for the ISP and sampled UQ dataset respectively. In each figure, the bar represents values up to 1 standard deviation above the median value, and the selected sample day is highlighted in blue colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are some variations in the distribution of features across different days of both datasets, we observe a fairly stationary behaviour, with the mean, median and standard deviation being largely consistent across the different days. Overall, we can conclude that the chosen sample days (day 25 and 33 for UQ and ISP datasets respectively) are a good representation of the overall dataset.

4.4. Feature scale-freeness

Table 4 provides additional details on the considered datasets, i.e., the size in terms of number of flows, the number of source and destination IP addresses, and the recording duration in seconds. The same applies to the remaining features listed in Table 3, such as the

Table 4
Details of considered datasets.

Dataset type	Dataset	Number of flows	Number of source IPs	Number of destination IPs	Recording duration (h)
Synthetic (**NetFlow**)	UNSW_NB15 (only benign)	1,507,090	36	30	137.25
	CIC_IDS2017 (only benign)	300,446	193	9701	8.1
	TON_IOT (only benign)	45,224	21	566	25.01
Real-World (**NetFlow**)	ISP [Single Day]	81,744,971	63,548	60,672	24.03
	UQ [Single Day]	47,911,959	269,282	129,212	24.02

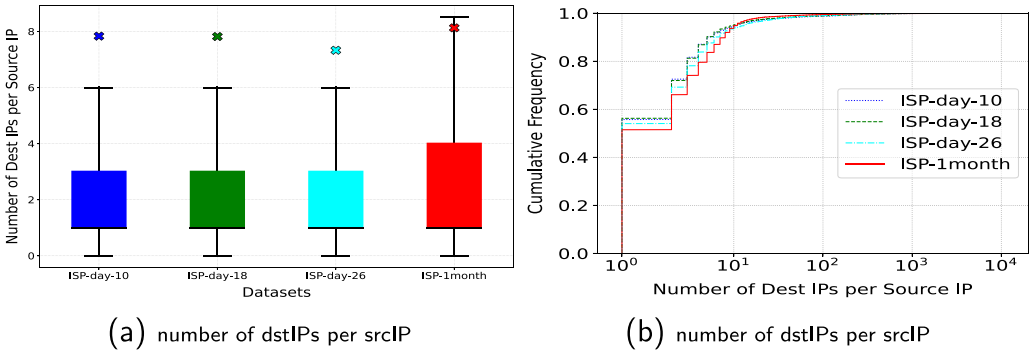


Fig. 3. Scale-freeness of the number of destination IPs per source IP in the ISP dataset. Figure (a) displays box plots, while Figure (b) shows cumulative distribution functions (CDFs). The figure compares three different single days with the one-month dataset.

average flow and packet size, which are statistical attributes or characteristics of network traffic. As can be seen, the scale of the real-world production networks (number of source and destination IP addresses) is significantly bigger than the typical lab-based networks used for the generation of synthetic datasets. For our analysis, we therefore need to ensure that the considered network traffic features are, as we refer to it, *scale-free*, i.e. they do not vary with the size of the network from which the data is collected, and consequently the volume of data that has been collected. If traffic features are not scale-free, we would expect their values to differ significantly between datasets of varying sizes, e.g. due to collection from networks of varying scales. Consequently, any observed statistical difference would have to, at least partially, be attributed to the network scale rather than the traffic characteristics.

The Flow and Packet Size primarily depend on network applications and protocols. Significant changes are not anticipated due to network scale. The next two features, Flow Duration and Packet Time, while can be influenced by factors like a network's geographical reach, applications used, and routing protocols, they are not directly affected by network scale alone. . This is not immediately obvious for the additional two features that we are considering in this paper, i.e. the *number of destination IPs per source IP*, and the *number of L7 protocols per source IP* (as shown in Table 3). We investigate the scale-free property of these two features in two aspects. First by considering datasets over different time scales, i.e. we compare a dataset for a single day with the full dataset, collected from the same network over the period of a month. We further investigated the scale-freeness across the network scale. The ISP dataset represents the aggregation of traffic records belonging to various ISP customers, which are identified by *VLAN* tags. In this analysis, the two above mentioned features are compared for a single customer and the entire ISP dataset. The selected customer dataset has 7135 source IP addresses, 21,292 destination IP addresses and the entire NetFlow data recording of one month equals to 5 GB, compared to 162 GB of the full dataset.

Fig. 3 shows the box plot and CDF for the *number of destination IP addresses per source IP*, for both three single day and the full 1-month version of our ISP dataset. We can see that the feature distribution does not vary significantly across different time scales of the dataset, and we can therefore assume it to be scale-free. The three selected days were chosen randomly to represent different time periods during data collection. Similar to Fig. 2, which explores flow size, flow duration, packet size, and packet time, there are no significant differences in the feature statistics across these various days. Fig. 4 shows the corresponding result for the feature of the *number of L7 protocols per source IP*. We can draw the same conclusion, i.e. that the feature exhibits the scale-free property.

Similarly, Figs. 5 and 6 show the comparison of the distributions of these two features for the network of a single customer and the entire ISP. In Fig. 5 the number of destination IP addresses per source IP are shown for both network scales. Fig. 6 shows the number of L7 protocols per source IP for both the customer network and the entire ISP network. As can be seen, in both cases the feature distributions for a single customer and the entire ISP network show a high level of similarity, which indicates a good degree of scale-freeness of the selected features in this respect.

The scale-free property was an important criterion for traffic features to be included in our analysis. Initially, we considered a much wider range of such features, but a lot of them failed the scale-freeness test. As an example, Fig. 7 shows the results for two such additional features: *number of L7 Protocols per dst Port*(a) and *number of source IPs per dst Port* (b) respectively. The figure shows the box plots of the feature distributions for the individual days, and also for the entire full-month dataset on the right (in red). We can clearly see that the distributions over the different time-scales vary significantly, and that these features are therefore not scale-free. To emphasise the distinction between these features and those utilised in our study, similar illustrations for the features listed in Table 3 have been created in Fig. 7-(c) to 7-(h). As observed, the one-month statistics closely mirror the single-day statistics.

5. Qualitative statistical analysis of datasets

The statistical analysis in this paper is divided into two main parts. First, we employ visual tools like box plots and cumulative distribution functions (CDF) to visually represent the similarities and differences between datasets originating from real-world and lab-based environments (synthetic datasets). This section offers a qualitative statistical analysis

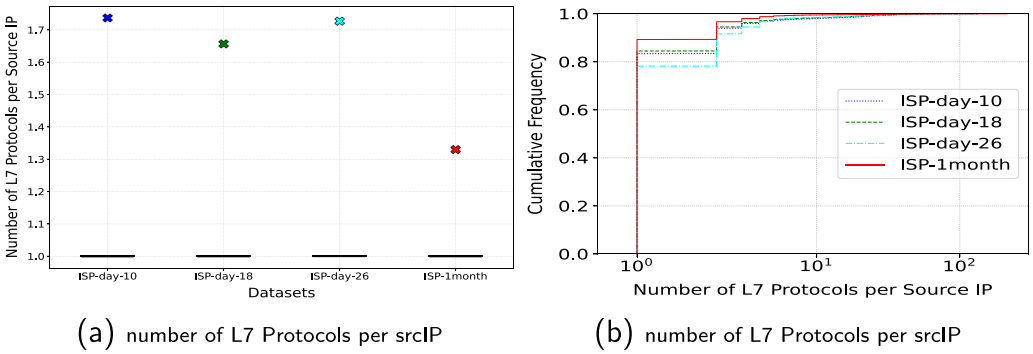


Fig. 4. Scale-freeness of the number of L7 protocols per srcIP in the ISP dataset. Figure (a) displays box plots, while Figure (b) shows cumulative distribution functions (CDFs). The figure compares three different single days with the one-month dataset.

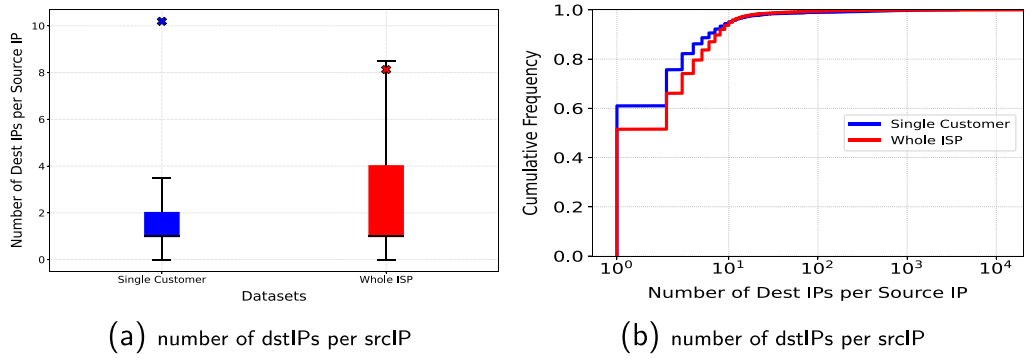


Fig. 5. Per customer and the whole ISP Scale-Freeness of the number of dstIPs per srcIP in ISP dataset (a) box plots, and (b) CDFs.

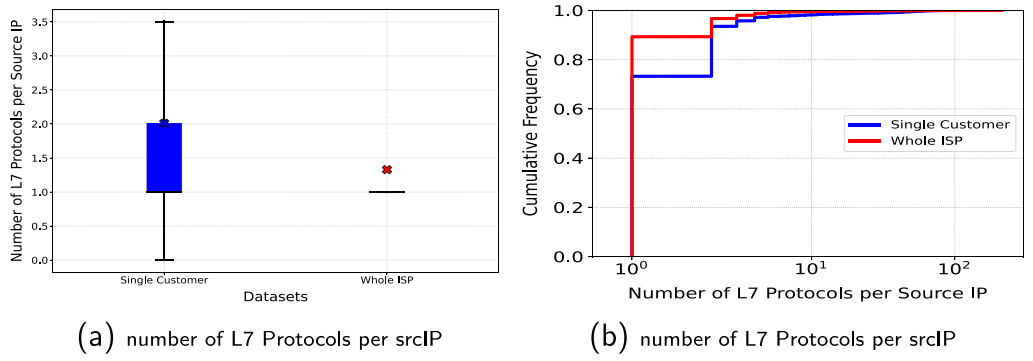


Fig. 6. Per customer and the whole ISP Scale-Freeness of the number of L7 protocols per srcIP in ISP dataset (a) box plots, and (b) CDFs.

of NIDS datasets, with a specific emphasis on benign traffic, focusing on the six selected traffic features outlined in Table 3.

Following the qualitative analysis, we delve into a quantitative statistical analysis of these similarities and differences in Section 6. To achieve this, we use the Wasserstein Distance metric, which stands as one of the most commonly employed methods for quantifying distances between distributions. This quantitative approach provides a deeper and more precise understanding of the statistical variations observed, enhancing the overall rigor of our analysis.

5.1. Flow duration

Flow duration has been identified as one of the key traffic characteristics by many previous works such as [8,11]. It has also been used for the network intrusion and anomaly detection in a wide range of studies, such as [30–32]. Fig. 8 shows the distribution of flow duration for the considered five datasets. In Fig. 8-a, the three left-most box plots correspond to the three synthetic datasets, and two right-most represent the two real-world datasets. In all box plots shown in this paper, the upper and lower *whiskers* indicate Quartile 1 – 1.5 IQR (Interquartile Range) and Quartile 3 + 1.5 IQR respectively. The outliers, i.e. values above the upper whisker and below the lower whisker, are not shown to increase clarity. The mean value is shown using a cross mark.

As can be seen, the flow duration distributions of the two real-world datasets (ISP and UQ) are quite similar and relatively long-tailed, as reflected both in their box plot and CDF representation. The three synthetic datasets have very similar distributions among themselves, with mostly very short flows, i.e. 94% to 99% of the flows have a sub-second flow duration. It is also evident that the mean values of ISP and UQ datasets have been positioned outside the distribution, a result attributed to the influence of outliers. In simpler terms, while flow durations are mostly very short for the majority of flows in both UQ and ISP datasets, causing the distributions to concentrate close to zero, the presence of exceptionally long flows pushes the mean values outside these distributions. We observe a very sharp contrast and discrepancy between the synthetic and real-world datasets, in regards to the flow

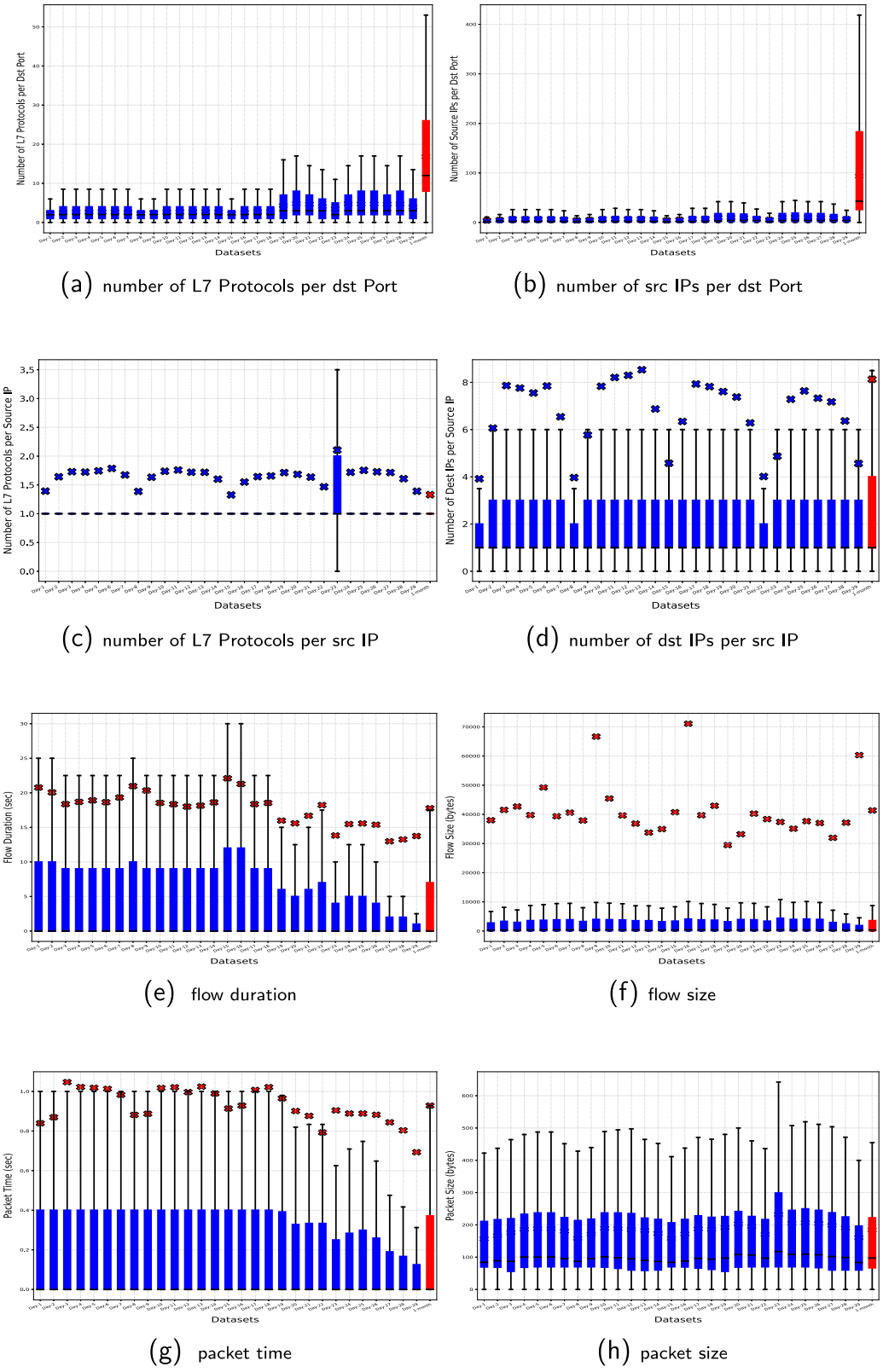


Fig. 7. Comparing Non-Scale-Free features displayed in (a) and (b) with Scale-Free features shown in figures (c) to (h) for the ISP dataset. In (a), the number of L7 protocols per destination IP is represented, while (b) shows the number of source IPs per destination Port. The subsequent figures depict (c) the number of L7 protocols per source IP, (d) the number of destination IPs per source IP, (e) flow duration, (f) flow size, (g) packet time, and (h) packet size.

duration of benign traffic. Given that flow duration is a critical traffic feature in the context of machine learning-based NIDS, we believe this is a significant observation.

To compare differences between real-world and synthetic datasets with those between the two real-world datasets, we have visualised the box plots representing the distribution of 3 days of each dataset

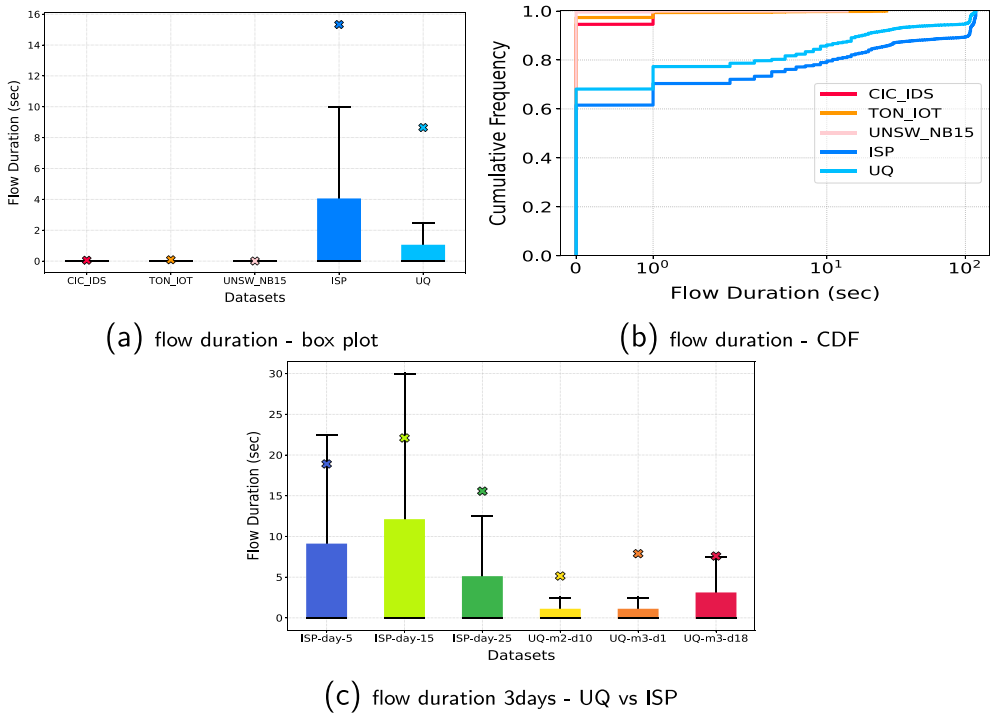


Fig. 8. The flow duration in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of flow duration in 3 days of UQ versus ISP.

in Fig. 8-(c). The differences between the distributions of the days of two real-world datasets and even between individual days within the same dataset are evident. However, these differences are less severe compared to those with the synthetic datasets visualised in Fig. 8-(b).

5.2. Flow size

The flow size (in Bytes)[2] is another important feature for both traffic characterisation and network intrusion and anomaly detection [33,34]. Fig. 9 displays the distribution of flow size in the five datasets in terms of their box plots and CDFs respectively. As was the case for flow duration, we observe that the two real-world datasets (ISP and UQ) show a close similarity in regards to their flow size distributions. This is interesting to note, given that the nature of these two networks are quite different, with one being a university network and the other an ISP with a large number of business customers of varying size.

In contrast, we see that the three synthetic datasets show widely differing distributions. TON_IOT is at the one end of the scale, with very small flows and a very narrow distribution and small IQR. At the other extreme, the UNSW_NB15 dataset has a very wide distribution and consequently a large IQR. Only the CIC_IDS dataset comes somewhat close to the real-world dataset, with a similar IQR. However, when we consider the CDF, we still see a significant difference to the UQ and ISP datasets, especially in the smaller flow size range.

We compared differences between real-world and synthetic datasets with those between the two real-world datasets by visualising box plots representing the distribution of 3 days of each real-world dataset in Fig. 9-(c). The differences between the distributions of the days of the two real-world datasets are noticeable, but less severe compared to those observed with the synthetic datasets UNSW_NB15 and TON_IOT in Fig. 9-(b).

As in the case of flow duration, we can conclude that the two real-world datasets exhibit a very similar distribution in terms of flow size. However, the synthetic datasets differ significantly among themselves and the real-world datasets, with the only minor exception of the CIC_IDS dataset.

5.3. Packet time

The average *packet time* is computed for each flow by dividing the flow duration by the total number of packets per flow, i.e. IN_PKTS + OUT_PKTS, as shown in row 3 of Table 3. Packet time in a computer network is generally made up of a number of different components, including transmission time, queuing time, processing time and propagation time. In our definition, packet time also considers flow idle time, during which no packet is transmitted, e.g. due to processing delays at the end hosts, and more importantly due to human-induced interaction delay. Packet time is therefore an important feature that not only captures the network characteristics such as link bandwidth, offered load and congestion, but also aspects of human-based traffic generation.

Fig. 10 shows the distribution of the (per-flow average) packet times for the five considered datasets We again see a similar pattern as for the previous traffic features. The two real-world datasets (ISP and UQ) show a very similar distribution, as reflected both in the box plot and CDF. The synthetic datasets also show a high degree of similarity among themselves, but their packet time distribution is vastly different from the one observed in the real-world datasets. This is also illustrated by the mean value, which is 2.28 Sec for ISP and 1.81 Sec for UQ, and 0.11, 0.22 and 0.02 Sec for the three synthetic datasets (CIC_IDS, TON_IOT and UNSW_NB15) respectively.

Like the previous features, we visualised box plots representing the distribution of 3 days of each dataset in Fig. 10-(c) to compare packet time differences between real-world and synthetic datasets with those

[2] We also considered flow size in terms of number of packets, but have not included the results in this paper, since the results are very similar in nature to the results in Bytes, as shown here.

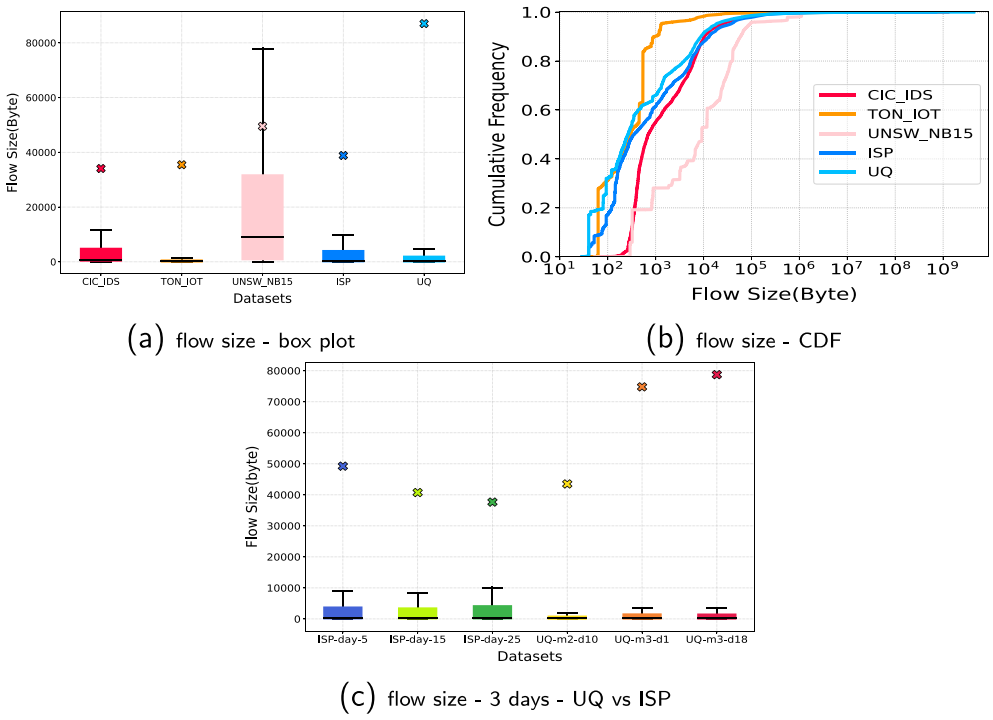


Fig. 9. The flow size (in bytes) in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of flow size for 3 days of UQ versus ISP.

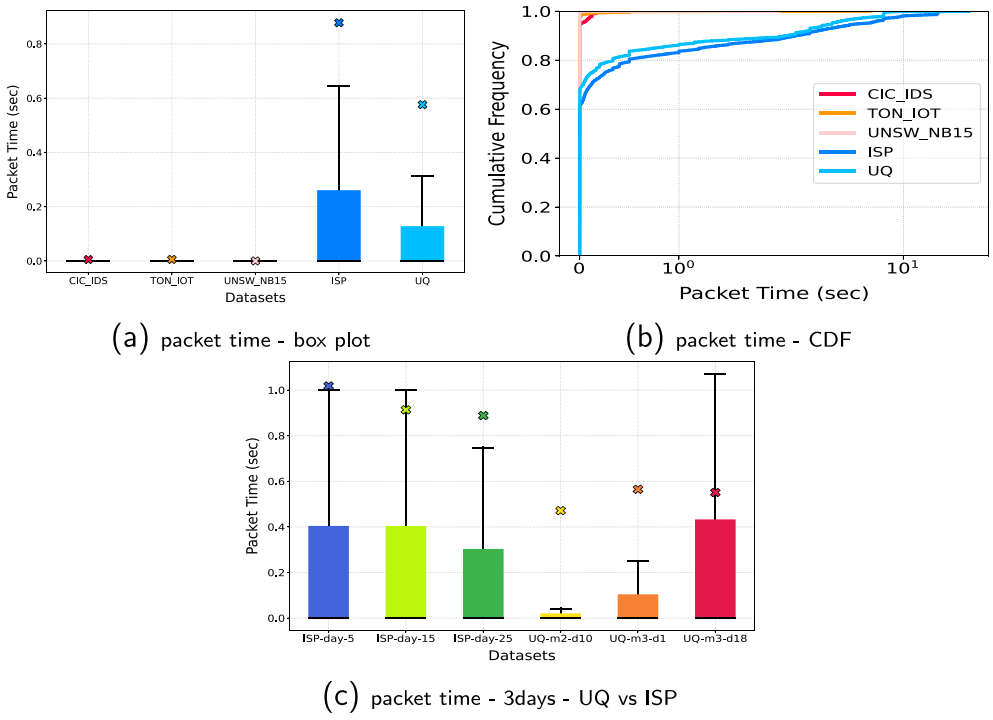


Fig. 10. The packet times in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of packet times in 3 days of UQ versus ISP.

between the two real-world datasets. The distribution of ISP days is very similar, but noticeable differences exist between the UQ dataset days. However, these differences are comparatively minor compared to the differences observed with all the synthetic datasets visualised in Fig. 10-(b).

5.4. Packet size

The (per-flow average) packet size is calculated by dividing the flow size in bytes by the total number of flow packets, i.e. IN_PKTS + OUT_PKTS, as shown in row 4 of Table 3. Packet size is an important

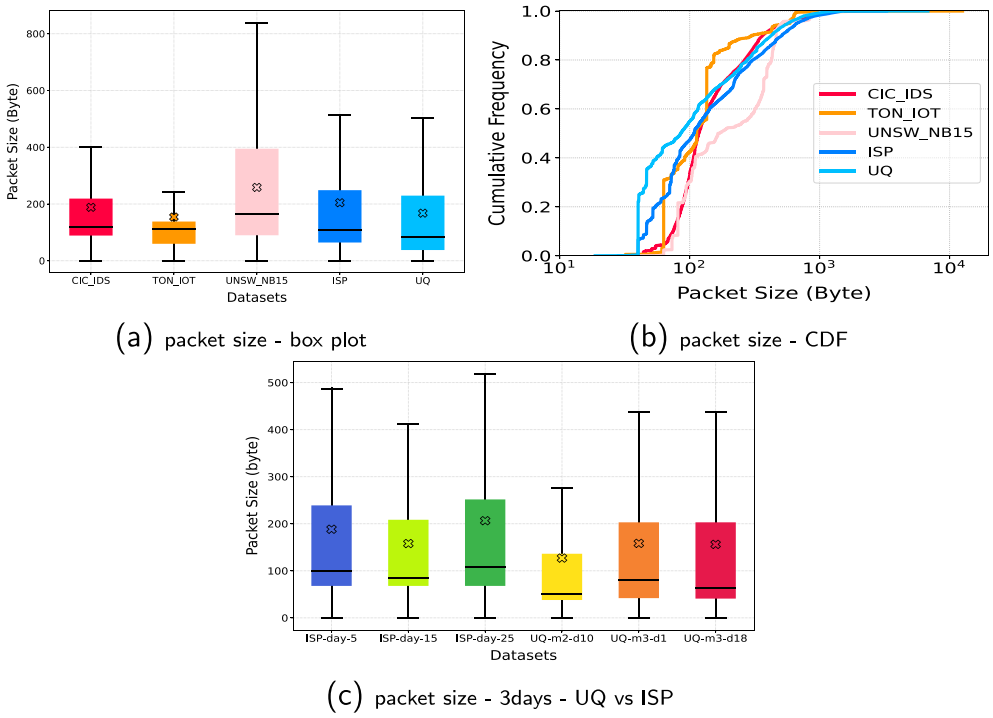


Fig. 11. The packet sizes in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of packet size for 3 days of UQ versus ISP.

traffic feature has been used widely in the context of NIDS, e.g. [34–36]. Fig. 11 shows the corresponding box plots and CDF for the five datasets. Our previously observed pattern is repeated here, with similar feature distributions for two real-world dataset, and with more widely varying distributions for the synthetic datasets. While the difference between the two groups is not as distinct as for the other features considered so far, we can still qualitatively observe a significant divergence, most clearly in the CDF representation. In particular, we notice for both the ISP and UQ datasets a higher representation of smaller packet sizes, with faster raise of the CDF, compared to all the synthetic datasets. In the middle range, with packet sizes of roughly 200 – 700 bytes, the synthetic datasets diverge strongly. Here, TON_IOT has a significantly higher representation (steep increase in CDF) than both real-world datasets, and UNSW_NB15 has significantly lower representation. In contrast, the CIC_IDS dataset relatively closely matches the CDF of the real-world datasets in that range.

We compared packet size differences between real-world and synthetic datasets, visualising box plots of 3 days for each real-world dataset in Fig. 11-(c). The two real-world datasets exhibit minor differences, notably smaller than those observed in Fig. 11-(b) with the synthetic datasets UNSW_NB15 and TON_IOT.

5.5. Number of destination IPs per source IP

The next feature we consider is the *number of destination IP addresses per a source IP*. This captures an important aspect of the traffic pattern (traffic matrix) in a network, i.e. the number destination hosts a source host is communicating with. The distribution of source and destination IP addresses has been investigated in a number of previous studies, in particular for the detection and identification of DDoS attacks [33,37,38], and hence we consider it relevant in the context of our analysis.

Fig. 12 shows the distribution of the feature in the usual box plot and CDF format. As before, we see a very close match between the distributions of the ISP and UQ dataset, and a notable variation among the three synthetic datasets. This is particularly evident from the CDF graph, with the CIC_IDS and UNSW_NB15 datasets furthest away from the two real-world datasets. Only the TON_IOT dataset seems to have a feature distribution that is somewhat close to the real-world datasets.

Similarly to previous features, we utilised box plots to depict the distribution of Number of Destination IPs per Source IP for 3 days of each real-world dataset in Fig. 12-(c). This comparison illuminates differences between real-world and synthetic datasets in contrast to differences between the two real-world datasets. The distribution of ISP and UQ days exhibits remarkable similarity, with closely aligned means and IQRs for most of the 6 days. This distinction becomes evident when compared to Fig. 12-(b), where differences between real-world and synthetic datasets are clearly visible.

5.6. Number of L7 protocols per source IP

The last feature investigated in this study, as shown in row 6 of Table 3, is the *number of L7 protocols per source IP*. This feature represents an important aspect of the network traffic pattern, since it considers the range of applications that end hosts are utilising, and we therefore believe it is important when considering characteristics of benign traffic in NIDS datasets.

Fig. 13 shows the distribution of this feature. Again, the pattern is repeated, with both real-world datasets showing a closely matched distribution, and the synthetic datasets varying widely among themselves, with only CIC_IDS coming somewhat close to the realistic datasets.

We examined the differences in the number of Destination IPs per Source IP between real-world and synthetic datasets, presenting the data using box plots for 3 days of each real-world dataset in Fig. 13-(c). In this visual representation, the daily statistics are remarkably similar, with minor discrepancies observed, notably on a single day from ISP. The differences between the two real-world datasets are significantly smaller than those depicted in Fig. 13-(b) with the synthetic datasets UNSW_NB15 and TON_IOT, although CIC_IDS is very close to real-world datasets.

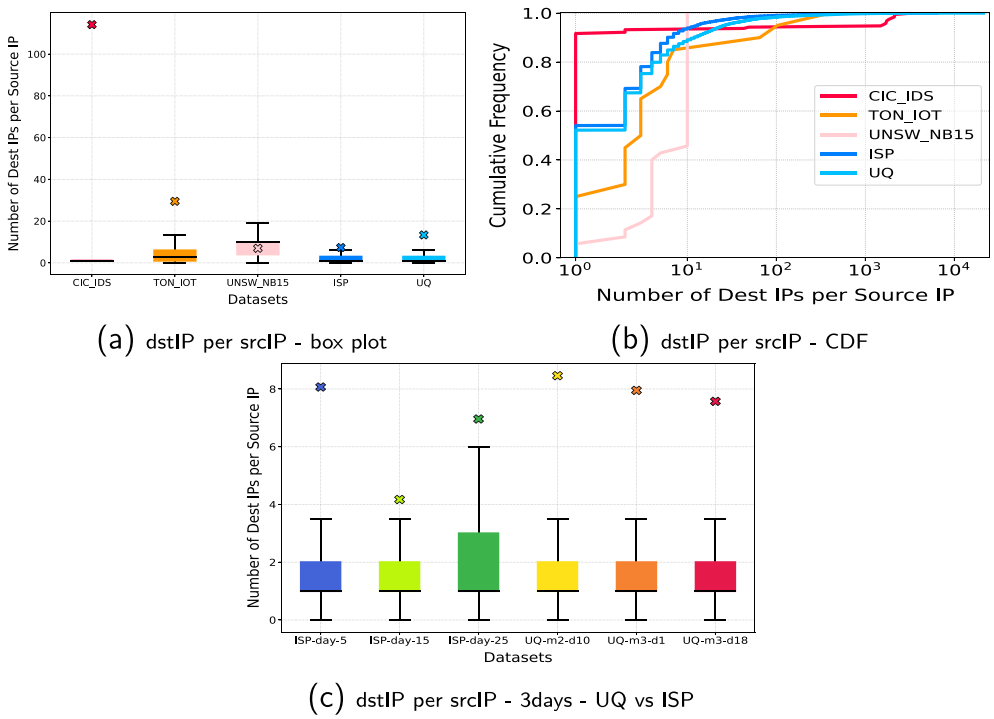


Fig. 12. Number of destination IPs per source IP in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of this metric in 3 days of UQ data versus ISP data.

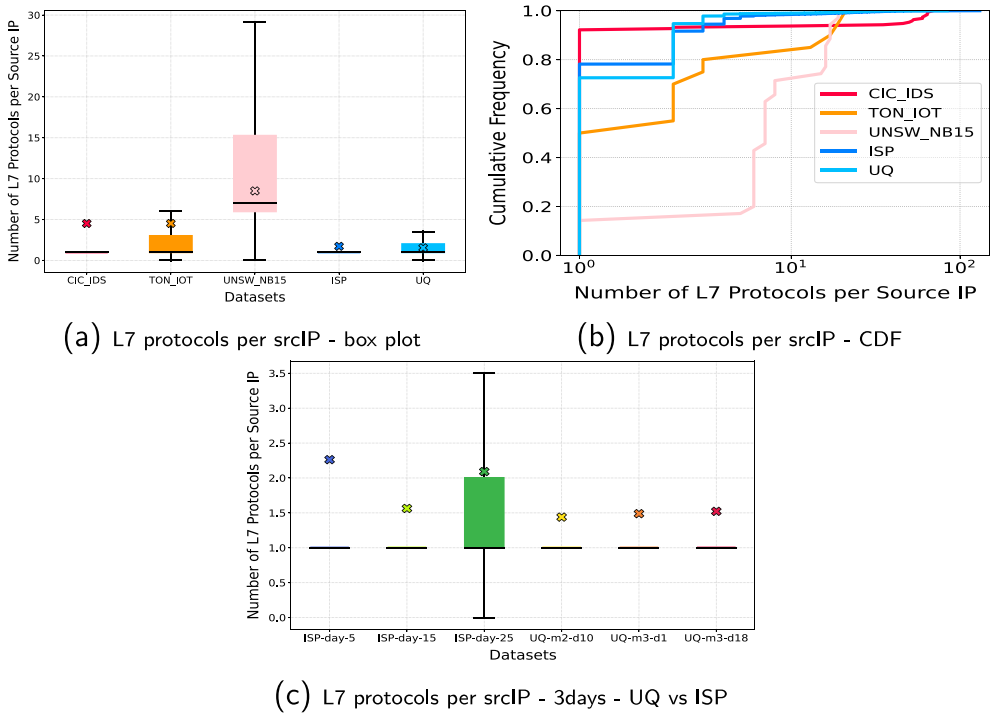


Fig. 13. Number of L7 protocols per source IP in five datasets (a) box plots, where the cross sign indicates the mean, and (b) CDFs. In (c) we are comparing the distribution of this metric in 3 days of UQ data versus ISP data.

In summary, our analysis of the considered traffic features, so far, shows a consistent pattern, across multiple features where the benign traffic of the synthetic datasets differs significantly from our two real-world datasets. While this difference is clear for some features, it requires further investigation for other features. To provide further depth to our analysis, the next section aims to quantify the differences between the various feature distributions using the Wasserstein metric.

6. Quantifying feature distribution distances

In the previous section, we visually presented the similarities and differences in feature distributions between the real-world and synthetic NIDS datasets using box plots and cumulative distribution functions (CDF). In this section, we employ a distance metric to precisely quantify these distinctions. Visual representation provided an initial

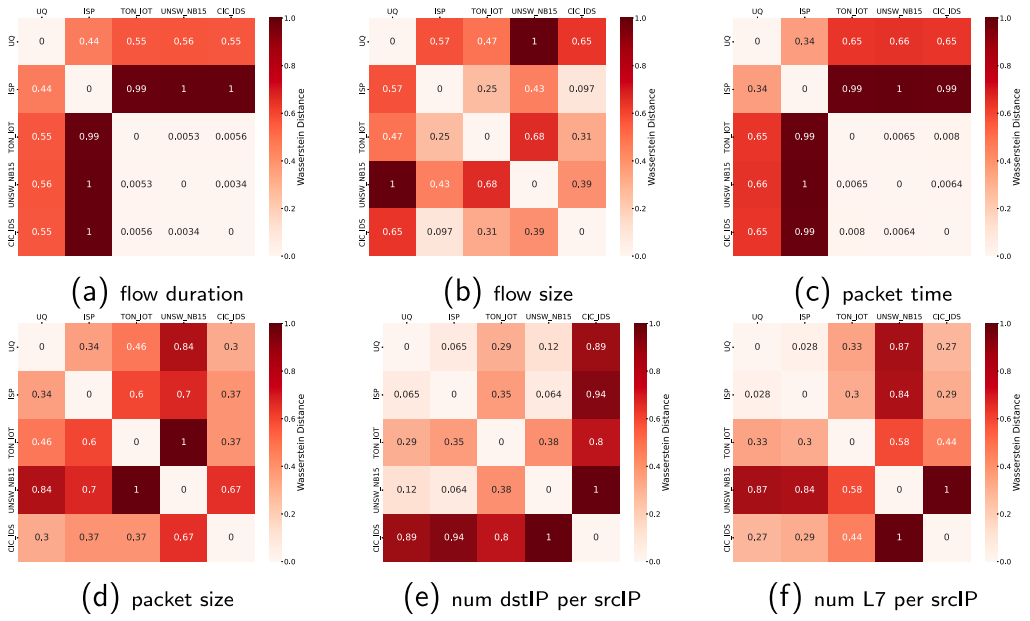


Fig. 14. Normalised Wasserstein distances of various feature distributions between the five datasets for (a) flow duration, (b) flow size, (c) (average) packet time, (d) packet size, (e) number of destination IPs per source IP, and (f) number of L7 protocols per source IP.

insight into the distinctions, and now we explore further by applying a quantitative approach to measure the extent of these differences.

While there are a range of potential metrics that can be used to measure the distance between two distributions/probability density functions, as reviewed in [39], we have chosen the (first) *Wasserstein distance*, also known as the *Earth Mover's distance*, for this purpose. A key advantage of the Wasserstein metric, in contrast to other commonly used methods to express differences between distributions, such as the Kullback–Leibler divergence, is that it is a proper metric, i.e. with the triangle inequality and symmetry properties. The discrete (first) Wasserstein distance $\mathcal{W}_1$ between the two distributions u and v is defined as follows [40]:

$$\mathcal{W}_1(u,v) = \inf_{\pi \in \Gamma(u,v)} \sum_{x,y} \|x-y\| \pi(x,y) \tag{3}$$

Here, $\|.\|$ represents the Euclidean norm and $\Gamma(u,v)$ the set of all distributions whose marginals are u and v on the first and second factors respectively. **inf** stands for *infimum*, i.e. the greatest lower bound. If S is a subset of set $\mathcal{T}$ (partially ordered), then **inf**(S) is the greatest element of $\mathcal{T}$ which is less than or equal to all elements of S. The first Wasserstein distance of u and v can also be stated in terms of their CDFs, U and V, as follows:

$$\mathcal{W}_1(u,v) = \sum_{-\infty}^{+\infty} |U - V| \tag{4}$$

This represents the absolute difference of the corresponding CDF curves [7]. This representation is more intuitive and suitable in our context, since in the previous section we have shown and the discussed the various feature distributions in terms of their CDFs.

Fig. 14 shows the normalised Wasserstein distances between each pair of the five datasets, in the form of a heat-map diagram, for all the six features studied in the previous section. Since the range of different features vary significantly, the raw Wasserstein distance values are normalised to a range of $[0,1]$ for each feature to allow inter-feature comparison. The correspondence between the Wasserstein distance values shown in Fig. 14 and the CDF curves shown in the previous section is clear. For instance, we clearly observe a 2x2 block in the top left corner of the heat maps, indicating the relative closeness of the two real-world datasets, in comparison to the synthetic datasets. While this separation is observable in all sub-figures, it is particularly evident in Figs. 14-e and 14-f. We also observe a distinct 3x3 block in the bottom right corner of some of the heatmaps, in particular Figs. 14-a and 14-c, showing the relative closeness of the three synthetic datasets in contrast to the real-world datasets. The outcomes of our comparisons do not depend on this specific arrangement. Even if the order of datasets were altered, the differences and similarities would persist and remain discernible. However, this intentional positioning was implemented to assist readers in comprehending the distinctions and similarities between the two groups of datasets. It served as a visual aid designed to enhance the reader's understanding.

In order to summarise the results shown in Fig. 14, we compute the mean Wasserstein distance values, averaged across the six considered features, shown in Fig. 15-a. The main takeaway from this table/heat-map is that the intra-group distances, i.e. the distances between instances of real-world datasets, and the distances between synthetic datasets, are relatively small. In contrast, the inter-group distances, i.e. the distances of real-world datasets and synthetic datasets, are comparatively larger.

To further illustrate this, Fig. 15-b shows the normalised mean Wasserstein distances of the five considered datasets, using the two real-world datasets as a reference. In the figure, the distance to the UQ dataset is represented on the x-axis, and the distance to the ISP dataset is represented on the y-axis. Consequently, as distance of UQ to itself is zero, and hence its data point is placed on the y-axis ($x = 0$), and similarly the ISP dataset is placed on the x-axis ($y = 0$). The clear conclusion from this figure is that the synthetic datasets differ significantly from the two considered real-world datasets in terms of the statistical feature distributions of the benign network flows.

7. Feature embedding

As a final approach to investigate the benign traffic of our considered NIDS datasets, we use feature embedding to map the network flow features into a 2-dimensional space for the purpose of visual inspection. For this, we consider the first four features in Table 3, i.e. *flow duration, flow size, packet time* and *packet size*.[3]

[3] *The number of destination IPs per source IP* and *the number of L7 protocols per source IP* were not included, since these values do not have a one-to-one correspondence with the flow records, and are computed over all dataset samples, as discussed in Section 4.1.

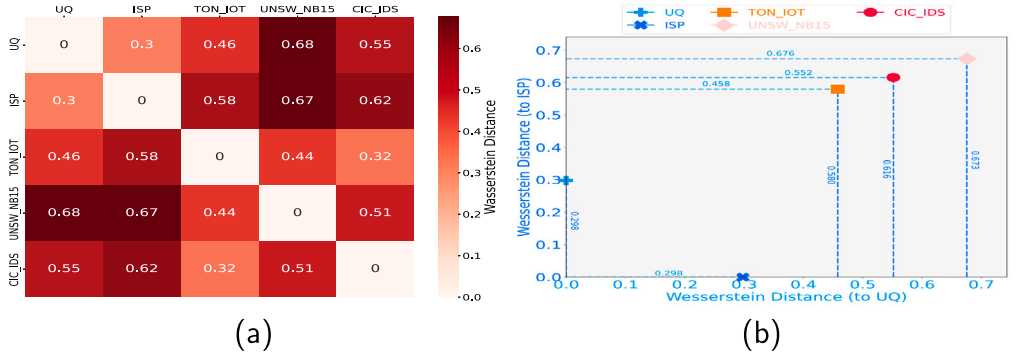


Fig. 15. Averaged normalised distances between datasets over all 6 features in Table 3 using (a) heatmap, and (b) distance to UQ and ISP.

The embedding techniques selected for our purpose are *Linear Discriminant Analysis (LDA)*, *Multi-dimensional Scaling(MDS)*, *Spectral Embedding* and *Principle Component Analysis (PCA)*. Since some of these methods are very computationally expensive, we used sampling to reduce the dataset size.

The resulting feature embeddings are plotted in Fig. 16. The horizontal axis represents the first embedded component and the vertical axis represents the second embedded component. As in the previous figures, the dark and light blue colours represent the UQ and ISP datasets, and the orange, light and dark red are used for UNSW_NB15, TON_IOT and CIC_IDS respectively. While the embeddings for the different algorithms look very different, we can observe a consistent pattern across the four subfigures. We see clear separation of the embedding data points of the real-world datasets (light and dark blue) from the embedding points of the synthetic datasets (orange, lignt and dark red). We also observe that the data points of the synthetic datasets are largely concentrated in a small, mostly 1-dimensional linear spaces, whereas the datapoints corresponding to the real-world datasets are much more spread out in the embedding space. This seems to indicate that the benign traffic in the synthetic datasets fails to reflect the range and variability of traffic patterns observed in real-world datasets. In summary, these qualitative findings from our feature embedding analysis are consistent with the results presented in earlier sections of this paper.

8. Discussion

The approach described in this paper involves using statistical characteristics extracted from network traffic, with a specific emphasis on NetFlow fields and features. These characteristics are used to investigate the similarities and differences between benign instances of real-world network traffic and three synthetic traffic datasets created in a controlled laboratory setting. However, it is crucial to recognise certain inherent limitations associated with this study.

Initially, a key point to consider is the numerical values calculated and demonstrated in this research. Our findings are derived from analysing two large sets of real-world samples, each with millions of records. However, it is vital to note that numerical values, such as Wasserstein distances, are specific to the particular features and datasets studied in this research. They may not be universally applicable across different datasets or features.

Secondly, it is important to acknowledge that the nature of network traffic is in constant flux due to the continuous introduction of new applications, protocols, and online games and services. This ever-changing landscape can directly impact the statistical measures computed and analysed within the framework of this study.

On another note, it is worth mentioning the impact of sampling on results visualised in this research. Various statistical visualisations and embeddings are employed in this study to clarify concepts and enhance presentations. Some of these visualisations and embeddings are generated from sub-samples of the main datasets. While every effort was made to conduct unbiased sampling, it is important to acknowledge that variations in sub-samples might lead to slight differences in the resulting visual representations.

Finally, consideration must be given to dataset distances, measured using Wasserstein distance across NetFlow features. These distances differ across various features; for instance, a synthetic dataset might be closer to a real-world dataset in one feature while farther in another. Take, for instance, UNSW_NB15's distances to UQ and ISP datasets in flow duration, which are 0.56 and 1 respectively, while in flow size, they are 1 and 0.43, as can be seen in Fig. 14-a and b. Consequently, dataset distance is not a uniform measure. Depending on the application, specific features may hold more significance. Then, the importance of a feature for an application determines the dataset's proximity to real-world data for that application.

Despite the limitations acknowledged, it is essential to highlight that our research has effectively identified noteworthy statistical distinctions between benign samples of real-world network traffic and synthetic datasets crafted within controlled environments. These differences span across several features frequently employed in network security analysis and applications. It is worth noting that, in many cases, applying the same distance measurement to real-world datasets yields considerably smaller divergences, underlining the unique challenges posed by synthetic datasets in accurately mimicking real-world complexities.

To ensure a fair and rigorous comparison with synthetic datasets, we excluded several features that exhibited **Scale-NOT-Free** behaviour. It is important to highlight that these exclusions were applied to a multitude of features, extending beyond the specific two features showcased in Fig. 7, namely, the *Number of L7 Protocols per src Port* and *Number of Src IPs per dst Port*. The decision to omit these features stemmed from the considerable differences in network scales between the smaller, controlled environments where synthetic datasets were generated and the larger, real-world networks.

Although these exclusions were made to maintain methodological integrity, it is essential to recognise that these features are prevalent in real-world networks. Their presence significantly influences the behaviour of applications and analyses relying on them. This makes the analysis results and application behaviour significantly divergent when applied to real-world data versus synthetic datasets. Consequently, ML-based NIDS algorithms and models trained using these synthetic datasets will exhibit varying behaviour when employed in real-world networks, especially when relying on these features.

9. Conclusion

High quality benchmark datasets are critical for the evaluation of machine learning models and systems. Due to the lack of access to current labelled datasets from real-world networks, research into ML-based Network Intrusion Detection Systems (NIDSs) has largely

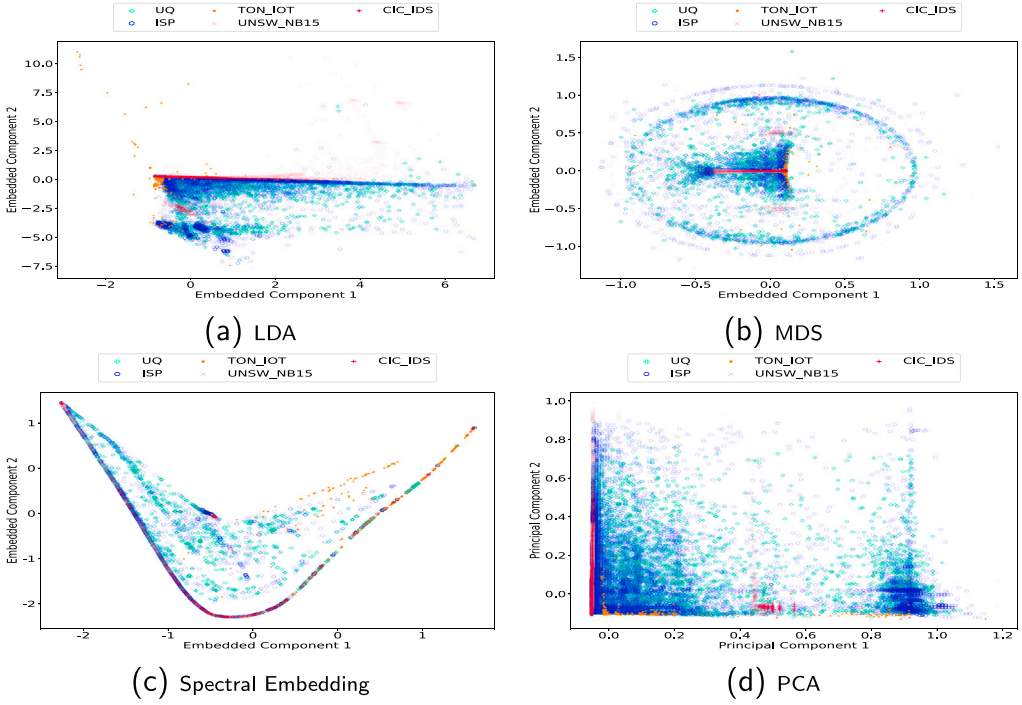


Fig. 16. Embedding of all datasets samples based on the first 4 features in Table 3 using (a) Linear Discriminant Analysis (LDA), (b) Multi-Dimensional Scaling(MDS), (c) Spectral Embedding, and (d) Principle Component Analysis (PCA).

relied on synthetic benchmark datasets. Using these datasets, excellent intrusion detection performance results have been published recently, approaching 100 percent across the key performance metrics such as Accuracy, F1 score, AUC, etc. However, we have not yet seen these excellent academic research results translated into near optimal NIDSs deployed in real-world production networks. The question of why that is the case provides a key motivation for our work presented in this paper. We were specifically interested to see if the statistical properties of synthetic NIDS datasets match those obtained from real-world networks. Due to the lack of access to labelled attack data from real-world networks, our investigation focuses on benign network traffic only.

The key contribution of this paper is the statistical analysis (and corresponding methodology) of traffic features of three widely used NIDS benchmark dataset, and their comparison with two real-world datasets. This work is enabled by converting the datasets to a common format, i.e. NetFlow. Our analysis based on six key traffic features showed that the two real-world datasets, obtained from very different production networks, have a high degree of similarity in terms of their feature distributions. We also observed that the three considered synthetic datasets exhibit largely similar feature distributions among themselves. The key finding of our analysis is the significant difference in feature distributions between the synthetic and real-world datasets. This raises the question of the generalisability of ML-models trained on synthetic NIDS datasets to real-world networks, and provides motivation for future research. We believe that by creating new NIDS datasets with more realistic background (benign) traffic, the translation of the excellent academic NIDS research results into real-world networks, and hence impact, can be significantly improved. In future work, we aim to extend our analysis to a wider range of datasets. We also aim to make our datasets publicly available. Towards this goal, we are working on a new improved anonymisation approach which will hopefully give network operators the confidence to share more of their network flow data with the research community.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Data is publicly available.

Acknowledgements

This research is made possible by an Advance Queensland Industry Research Fellowship, Australia, grant number RM2019002409.

References

[1] University of California, Irvine. KDD cup 1999 data. 1999, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Accessed 30 July 2020].

[2] Moustafa N, Slay J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: Military communications and information systems conference. IEEE; 2015, p. 1–6.

[3] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSP 2018 - Proceedings of the 4th international conference on information systems security and privacy, vol. 2018-janua 2018;108–16.

[4] Moustafa N. New generations of internet of things datasets for cybersecurity applications based machine learning: Ton_IoT datasets. In: Proceedings of the EResearch Australasia conference. 2019, p. 21–5.

[5] Sarhan M, Layeghy S, Moustafa N, Portmann M. NetFlow datasets for machine learning-based network intrusion detection systems. In: Big data technologies and applications. Cham: Springer International Publishing; 2021, p. 117–35.

[6] Sarhan M, Layeghy S, Moustafa N, Portmann M. Towards a standard feature set of NIDS datasets. 2021.

[7] Ramdas A, Trillos NG, Cuturi M. On wasserstein two-sample testing and related families of nonparametric tests. Entropy 2017;19(2):47.

[8] Kevin Thompson, Gregory JMiller, Rick Wilder. Wide-area internet traffic patterns and characteristics. IEEE Netw 1997;11(6):10–23.

[9] Anukool Lakhina, Konstantina Papagiannaki, Mark Crovella, Christophe Diot, Eric DKolaczyk, Nina Taft. Structural analysis of network traffic flows. SIGMETRICS Perform Eval Rev 2004;32(1):61–72.

[10] Liu L, Jiang HH, Rui WZ, Wang J. Study on the characteristics of network traffic based on STFT. In: Proceedings - 2015 2nd international conference on information science and control engineering. IEEE; 2015, p. 485–8.

[11] Theophilus Benson, Aditya Akella, David AMaltz. Network traffic characteristics of data centers in the wild. In: Proceedings of the 10th ACM SIGCOMM conference on internet measurement. ACM; 2010, p. 267–80.

[12] Kandula S, Sengupta S, Greenberg A, Patel P, Chaiken R. The nature of datacenter traffic: Measurements & analysis. In: Proceedings of the ACM SIGCOMM internet measurement conference. 2009, p. 202–8.
[13] Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions. SIGCOMM Comput Commun Rev 2005;35(4):217–28.
[14] Soule A, Ringberg H, Silveira F, Rexford J, Diot C. Detectability of traffic anomalies in two adjacent networks. In: International conference on passive and active network measurement, vol. 4427 LNCS. Springer Berlin Heidelberg; 2007, p. 22–31.
[15] Andreas Kind, Marc PhStoecklin, Xenofontas Dimitropoulos. Histogram-based traffic anomaly detection. IEEE Trans Netw Serv Manag 2009;6(2):110–21.
[16] Mchugh J. Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Trans Inf Syst Secur 2000;3(4):262–94.
[17] Mahoney MV, Chan PK. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. Lecture Notes in Comput Sci 2003;2820(Ll):220–37.
[18] 1999 DARPA Intrusion Detection Evaluation Dataset | MIT Lincoln Laboratory.
[19] Thomas C, Sharma V, Balakrishnan N. Usefulness of DARPA dataset for intrusion detection system evaluation. Data Min Intrusion Detect Inf Assur Data Netw Secur 2008 2008;6973(March 2008):69730G.
[20] Roesch M. Snort – lightweight intrusion detection for networks. In: Proceedings of LISA 99: 13th systems administration conference. 1999, p. 229–38.
[21] MIT Lincoln Laboratory: DARPA intrusion detection evaluation. 2021.
[22] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications. 2009, p. 1–6.
[23] 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. Inf Secur J 2016;25(1–3):18–31.
[24] Gharib A, Sharafaldin I, Lashkari AH, Ghorbani AA. An Evaluation Framework for Intrusion Detection Dataset. In: ICISS 2016 - International conference on information science and security 2016, no. cic. IEEE; 2017, p. 0–4.
[25] Dwibedi S, Pujari M, Sun W. A comparative study on contemporary intrusion detection datasets for machine learning research. In: Proceedings - 2020 IEEE international conference on intelligence and security informatics. 2020.
[26] Kilincer IF, Ertam F, Sengur A. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Comput Netw 2021;107840.
[27] Ring M, Wunderlich S, Scheuring D, Landes D, Hotho A. A survey of network-based intrusion detection data sets. Comput Secur 2019;86:147–67.
[28] Ntop. nProbe, an extensible NetFlow v5/v9/IPFIX probe for IPv4/v6, no. july. 2017, p. 1–60.
[29] Sivanathan A, Gharakheili HH, Loi F, Radford A, Wijenayake C, Vishwanath A, et al. Classifying IoT devices in smart environments using network traffic characteristics. IEEE Trans Mob Comput 2019;18(8):1745–59.
[30] Aleksey AGaltsev, Andrei MSukhov, Balandin S, Koucheryavy Y, Hu H, Sukhov AM. Network attack detection at flow level. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011, p. 326–34.
[31] Shahriar MH, Haque NI, Rahman MA, Alonso M. G-IDS: Generative adversarial networks assisted intrusion detection system. 2020, arXiv.
[32] Rigaki M, Garcia S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. In: Proceedings - 2018 IEEE symposium on security and privacy workshops. IEEE; 2018, p. 70–5.
[33] Gilberto Fernandes, Luiz FCarvalho, Joel JPCRodrigues, Mario Lemes Proença, Fernandes G, Carvalho LF, Rodrigues JJPC, Lemes M, Jr Proença. Network anomaly detection using IP flows with principal component analysis and ant colony optimization. J Netw Comput Appl 2016;64:1–11.
[34] Wang W, Sheng Y, Wang J, Zeng X, Ye X, Huang Y, et al. HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. IEEE Access 2017;6:1792–806.
[35] Duffield N, Haffner P, Krishnamurthy B, Ringberg H, Nick Duffield, Patrick Haffner, et al. Rule-based anomaly detection on IP flows. In: IEEE INFOCOM 2009, vol. 08544. IEEE; 2009, p. 424–32,.
[36] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, Asaf Shabtai, Mirsky Y, Doitshman T, et al. Kitsune: an ensemble of autoencoders for online network intrusion detection, no. February. 2018, p. 18–21.
[37] Jieren Cheng, Jianping Yin, Yun Liu, Zhiping Cai, Min Li, Cheng J, et al. DDoS attack detection algorithm using IP address features. In: International workshop on frontiers in algorithmics. Berlin, Heidelberg: Springer; 2009, p. 207–15.
[38] Anna LBuczak, Erhan Guven, Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor 2016;18(2):1153–76.
[39] Cha S-H. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Methods Appl Sci 2007;1(4):300–7.
[40] Herrmann V. Wasserstein GAN and the Kantorovich-Rubinstein duality - Vincent Herrmann. 2017.