Electronic Preprint for Journal of Information Processing Vol.29

Regular Paper

Novel Bi-directional Flow-based Traffic Generation Framework for IDS Evaluation and Exploratory Data Analysis

Korakoch Wilailux[1,a)]   Sudsanguan Ngamsuriyaroj[1,b)]

Received: June 9, 2020, Accepted: December 1, 2020

Abstract: Flow-based network traffic information has been recently used to detect malicious intrusion. However, several available public flow-based datasets are unidirectional, and bidirectional flow-based datasets are rarely available. In this paper, a novel framework to generate bidirectional flow-based datasets for IDS evaluation is proposed. The generated dataset has the mixed combination of normal background traffic and attack traffic. The background traffic is based on the key traffic feature of the MAWI network traffic traces, and five popular attack traffics are generated based on their statistical traffic features. The generated dataset is characterized using the PCA approach, and we found out that benign and malicious traffic are distinct. With the proposed framework, a dataset of bi-directional flow-based traffic is generated and it would be used for evaluating an effective intrusion detection engine.

Keywords: traffic modeling, protocol behavior model, traffic generation, network intrusion detection

1. Introduction

Several datasets for Intrusion detection system (IDS) evaluation are publicly available such as DARPA99 [37], KDD99 [5], and DEFCON [4], [36]. However, these datasets are outdated and developed to evaluate packet-based network intrusion detection systems such as Snort [3], Suricata [6], and Bro [39], which recognize malicious attempts using deep packet inspection signature matching. The problems of packet-based inspection have been reported toward recognizing network attacks in contemporary network [9], [22], [55]. Therefore, flow-based inspection techniques are proposed as a complementary method for secure network infrastructure. Recently, many researchers focus on flow-based analysis techniques, which classify attacks by analyzing the aggregated information extracted from connection or session information [27], [53], [55]. However, research communities are showing the demand for good flow-based datasets, representing realistic network characteristics and contemporary attacks [43], [48], [50], [51].

There are techniques proposed to generate a dataset for evaluating the intrusion detection system performance [14], [24], [28]. A researcher might replay captured traffic, performs traffic spawning using statistical analysis of network traces, simulates source-based traffic, or captures live-traffic [28], [38], [45], [48], [52]. There are also other methods to generate the datasets, either by capturing traffic from a closed network or by synthesizing traffic using a simulation tool [10], [28], [53]. Capturing then labeling live traffic would give a better representation of real attacks, but it could also raise a problem related to privacy issues because there is a risk of exposing personal user information. Also, it could bring incomplete annotation, since the traffic may contain unknown attacks [43], [50], [51]. If a dataset is not accurately tagged, it will depreciate a dataset utility for evaluating the IDS. Therefore traffic simulation is an analogous method that is more difficult to perform but fits the potential for more flexibility in creating a realistic and completely labeled dataset.

Traffic simulation techniques have been extensively applied to create the IDS evaluation dataset [38], [44], [45], [46], [54]. Some researches to date focus on modeling unidirectional flow-based traffic features, therefore many of the up-to-date datasets lack either substantial session characteristics of bi-directional flow features or contemporary attack tactics. This research aimed to replicate realistic bi-directional background traffic as well as implanting the modern attack during the simulation. In this study, a bi-directional traffic simulation framework is proposed to generate a bi-directional flow-based dataset for evaluating the flow-based intrusion detection system. The framework is designed to replicate session traffic features embedded in the bi-directional network flow using a descriptive traffic model, and also enables adding a broad range of attacks generally observed in the present network. The following structural bi-directional flow-based features are modeled and used in the framework – flow size, flow duration, and packet count per second. The framework can generate a bi-directional flow-based dataset and the generated structural traffic features are tested against the real network data to verify that the characteristics are statistically similar to the real network trace.

We conducted the experiments to ascertain whether the proposed framework could generate similar bi-directional traffic

[1] Faculty of Information and Communication Technology, Mahidol University, Nakornpathom, Thailand
[a)] korakoch.w@navy.mi.th
[b)] sudsanguan.nga@mahidol.ac.th

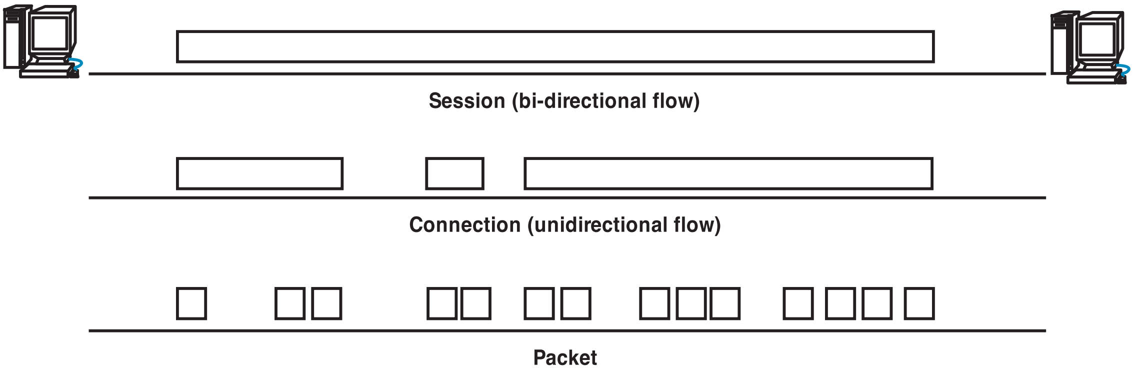
Session (bi-directional flow)
Connection (unidirectional flow)
Packet

Fig. 1 Session behavior captured in bi-directional traffic flow.

flows to the original bi-directional traffic flow. The MAWI [20] traffic traces are selected as the original traffic because they are traffic repositories containing comprehensive information at the Internet backbone traffic, which are also publicly accessible and redistributable traffic data sets. In the experiments, the traffic was modeled as a stationary process and its structural session traffics of DNS, FTP, HTTP, SMTP, and SSH were observed as the descriptive parameters, then used in benign traffic creation. The following malicious traffic was generated – network scanning, host scanning, web vulnerability scanning, SSH brute-forcing, and SQL injection. The statistical similarity of the simulate dataset to the original MAWI network traffic was measured by comparing the empirical cumulative distribution function (ECDF) to validate the realism of the simulated dataset. The created traffic features are statistically similar to the original traffic features. The exploratory data analysis results also confirm the usability in evaluating an intrusion detection system, because the experimental results have shown the distinction of the flow's features between benign traffic and various types of attacks.

In this paper, our contributions are listed as follows. Firstly, we present a novel framework that employs bi-directional flow traffic models to generate a dataset for the bi-directional flow-based intrusion detection system. Secondly, we deliver the descriptive bi-directional traffic models for realistic background traffic simulation. Lastly, the sample dataset created through the framework is exploratorily analyzed to validate its realism and utilization in evaluating a flow-based intrusion detection system.

The structure of the paper is presented as follows. Section 2 summarizes and discusses the related work on dataset generation and exploratory data analysis. Section 3 introduces our works. Section 4 describes the generated dataset and presents its exploratory data analysis results. Section 5 draws the conclusion and offers discussion on the proposed work.

2. Background and Literature Reviews

Network traffic usually consists of two-way communication transmitted over transmission control protocol – a client and a server, occurring in the amount of traffic flowing between ingress *i* and egress *j* is not independent of the amount of traffic streaming from ingress *j* to egress *i*. Researchers [41], [49] practice that assumption to model the aggregated-traffic, assuming the independence assumption of packets to estimate Origin-Destination (OD) flow count between end-hosts. However, Erramilli et al. [18] argued that the Internet traffic should not hold the packet's independence assumption because an aggregated traffic should be concerned as collections of connections, causing the bidirectional nature of a communication session. For example, web traffic tends to have a greater volume of traffic moving in the reverse direction than in the forward direction, while P2P traffic may show higher symmetry. They also explained that the simpler forms of traffic models were adequate for applications such as imitating application characteristics.

There are two methods to describe network traffic – the analytical and simulation approaches [30]. The analytical approach describes traffic mathematically using applied mathematics tools such as queuing and the probability theory. Usually, traffic modeling involves the definition of analytic models, which are used to compare and analyze the empirical model of traffic. Traffic modelers create models using traffic measurement data to represent the characteristic of the application protocol.

Generally, network traffic can be characterized at different levels: packet, flow, and session [15]. **Figure 1** depicts the general concept on different type of network traffics. Packet-Based modeling focuses on the arrival process of packets, ignoring the interaction nature persisting in the traffic. Flow-Based modeling focuses on analyzing consecutively transmitted packets in one direction of an exact pair of endpoints. A uni-directional flow is defined such that the flow is composed only of packets sent from a single endpoint to another single endpoint. As declared in RFC5103 [1], a bi-directional flow is a bi-directional measure for a network parameter by metering or accounting for certain attributes in the network session composed of packets sent in both directions between two endpoints, which is potentially associated with human activity. This association provides additional data that can be useful for security analysis tasks [12].

2.1 Traffic Modeling and Its application in Dataset Generation

Typically, a packet-level network trace is captured and analyzed to construct a simulated traffic model using some random variables such as packet inter-arrival time and packet size. Aikat et al. [8] have presented that the dataset generated from various types of TCP applications showed significantly different network characteristics. They also reported that the round trip time creates a small impact on the performance, while the structure of the TCP application workload appoints a great impact. Thus, the details of traffic measurements are very important if it was desired to regenerate credible simulated traffic. Gomes et al. [26] also have modeled application traffics of traffic sources, derived from real traffic traces. The Web, VoIP, streaming, and file-sharing traffic have been studied using distribution fitting and auto correlation techniques. The Kolmogorov-Simirnov was used to measure the goodness of fit to the theoretical distribution. In their work, the packet size, inter-arrival times and byte count were investigated. Floyd et al. [19] have shown that application protocol behavior at packet-level was insufficiently represented the contemporary network traffic since the congestion control mechanism was hardly modeled using packet-based approaches.

Golaup et al. [25] have proposed a framework for simulating multimedia traffic using a source model, which represented a set of multimedia traffic component; consisting of a block, transaction, and streaming characteristic. Each traffic component was statistically analyzed and combined into a source model. The behavior of multimedia traffic was mimicked using On-Off modeling techniques and the multimedia traffic in HTTP session was identified when no silence period was found in an active session. The byte count during the On-Off period was analyzed and used in traffic recreation. Hoflack et al. [29] have also presented a session-based web server traffic model, extending the train arrival process. They assumed that a user generates non-interrupted connections until the session ends, where each session must contained at least one packet per session. The buffer occupancy and packet delay were investigated and presented to derive the mean session delay under the first-come-first-served queuing model for a packet. The model was evaluated with an actual trace of web server traffic. The finding suggests that an approximation of HTTP session byte should be done by heavy-tailed distribution.

Barakat et al. [11] have applied a Poisson shot-noise process to simulate the traffic at the Internet backbone. The unidirectional flow arrival rate, flow size, and flow duration were characterized. Their traffic model also approximated the average traffic on the Internet backbone at short timescales. In addition, as supported by the empirical data, they concluded that traces with less than 30-minute duration tended to have the arrival process stationary. The finding allowed Afanasiev et al. [7] to examine link utilization and the number of active flows, which were used in characterizing the state of any network. The correlation between link utilization and the number of active flows was visualized as measuring the network quality and to discover possible bottlenecks on the Internet backbone.

Kim et al. [32] have also analyzed the general characteristic of application traffic at connection-level using unidirectional flow. The traffic statistics such as connection duration, packet count, and bytes count were statistically analyzed and presented. They defined two types of connection characteristics – traffic without corresponding communication and traffic with reverse-corresponding communications, while Lee et al. [34] had comprehensively reviewed several stochastic processes used in modeling network traffic. The traffic rate measured in packets per second and the packet inter-arrival time was modeled using the fractional Gaussian noise, the linear fractional stable motion, Poisson process, and conservative binomial cascades. Recently, Wette et al. [56], had characterized connection-level traffic in data centers using probability distribution functions and applied it on generating traffic for a data center. The amount of data in bytes that was transmitted between end hosts, where the flow-size distribution of payload, the flow-size distributions of ACK, and the combination of payload and ACK flow were distinguished. The total amount of traffic for a given period is estimated using the distribution of flow inter-arrival time. The statistical properties of simulated and observed traces were compared using Quantile-Quantile plots (Q-Q plots) rather than measuring the goodness of fit.

The arrival process of HTTP, FTP, POP3, SSH, and SMTP were modeled and reported by Song et al. [47]. Their works focused on deriving statistical models of traffic at connection-level. Some random variables of application traffic such as HTTP request arrivals, bytes transferred were modeled. However, other significant traffic parameters that describe the application behavior such as connection duration and packets count per connection were not modeled and reported. The inter-mixing traffic modeling techniques were applied by Song et al. [48], as well. The application of Markov model in capturing user-activity using Netflow-like statistic was shown. To estimate the characteristic of an application protocol, the connection traffic structure was estimated using Expectation-Maximization (EM) and mixture model technique. The host-based behavior of the connections was presented. However, the authors have not presented the attack simulation method in their work. Yang et al. [57] have also worked at connection-level modeling. In their work, the use of active unidirectional flow as key simulation object to mimic the characteristic of the Internet traffic was investigated. The structural model of flow arrival rate, the active flow, and its duration had been modeled and evaluated using 600 seconds-long Abilene traces. Their work assumes an assumption that all application protocols had similar traffic structure. We argued that different protocols may have variant behaviors. Our work extends the use of structural models and payload metrics.

Vishwanath [54] have reported their design and implementation of a traffic generator software named Swing. Their work centered on reproducing structural Intra-net traffic using models extracted from packet header traces. They extracted session models using unidirectional flow information and two specific thresholds. Their framework created realistic application workloads mimicking the presence of traces at the Internet. However, their effort focused on regenerating trace without any network attack. Sarraute et al. [42] have presented a network simulator, focusing on generating the attack traffics. The framework was implemented using proxy system calls and multi-platform agents, which allow remote system call execution on the vulnerable host. The attacks were modeled as probability actions, which were executed by autonomous agents. Sommers et al. [46] have published a tool to generate a network-wide flow record, and showed its capability to evaluate anomaly detection techniques. The tool could generate realistic unidirectional flow records and some types of anomalous flows. They compared traffic volume generated by tools with one created by ns2-simulator, showing the capability of recreating traffics with known characteristic. However, the tools could

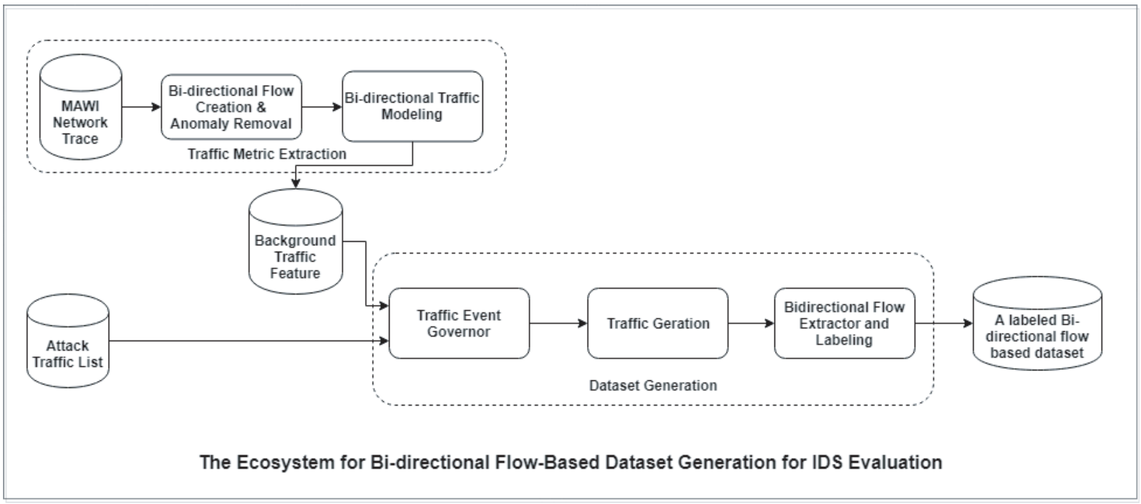
MAWI Network Trace
Bi-directional Flow Creation & Anomaly Removal
Bi-directional Traffic Modeling
Traffic Metric Extraction
Background Traffic Feature
Attack Traffic List
Traffic Event Governor
Traffic Geration
Bidirectional Flow Extractor and Labeling
A labeled Bi-directional flow based dataset
Dataset Generation
The Ecosystem for Bi-directional Flow-Based Dataset Generation for IDS Evaluation

Fig. 2 The ecosystem for bi-directional flow-based dataset generation for IDS evaluation.

not generate semantic attacks, i.e., brute-forcing the application protocols.

Shiravi et al. [45] have proposed a framework for generating a benchmark dataset through the use of profiles, representing events and statistical properties extracted from a network trace. The background traffic was created by extracting statistic parameters at connection levels in a laboratory network and used in regenerate benign traffic without structural characteristics. There were two types of profiles – α and β profile. The α profiles described attacks scenarios and their relevant parameters, while β profiles explained syntactic background traffic, modeled at connection-level. They have demonstrated the usefulness of their framework. However, their key idea focused around simulating the behavior of users but our work was designed to simulate the characteristic of some important application protocols. Maciá-Fernández et al. [38] also presented a dataset for a cyclostationarity-based intrusion detection system, offering time-span dataset. However, their work was proposed and developed based on the unidirectional flow concept and limited the type of network attack to the cyclostationarity-related attack such as SSH scan attack and SPAM attack. Elejla et al. [17] also presented a dataset for ICMPv6 DDoS attack detection. The dataset was created and modeled using unidirectional flow features, which replicated the campus network traffic. The dataset was designed specifically for the IPv6 network management attack, therefore other important network attacks were left out.

Recently, Generative Adversarial Networks (GAN) gained more interest in generating network traffic. Dowoo et al. [16] proposed PcapGAN, Packet Capture File Creator, which applied a style-based Generative Adversarial Networks to generate pcap network trace. The generated packet data could be accessed and analyzed by general network analysis software such as Wireshark. The authors also converted the generated packet-based traffic into the unidirectional flow using KDD99 extractor and quantified the similarity of the original dataset and the generated data through the analysis of the principal component distribution of the two datasets. Ring et al. [40] also employed GAN to generate flow-based network traffic. The authors introduced Embedding-based Improved Wasserstein Generative Adversarial Network (E-WGAN-GP) to synthesize unidirectional flow attributes, which were generated based on the probability distribution of each traffic attribute and was pulled independently from other features. Their experimental results showed that the proposed GAN-based approach regenerated a realistic unidirectional flow dataset. However, attack traffic generation was excluded from their investigations.

3. Bi-directional Flow-based Dataset Generation

It is crucial to generate a realistic dataset, which resembles and acts as realistically as possible. We decided to regenerate background traffics using the bi-directional flow concept, since modeling at bi-directional flow level further allows us to emulate human user's behavior, such as streaming and typing in a single flow. **Figure 2** depicts the ecosystem and our proposed framework to generate a bi-directional flow dataset for IDS evaluation, which is composed of two main components – Traffic Extraction and Modeling, and dataset generation.

3.1 Traffic Extraction and Modeling

We defined a bi-directional flow as a sequence of packets belonging to the same session. Also, a flow was created whenever there is an active timeout (30 minutes), or inactive timeout (15 seconds), or when FIN or RST flag was produced. A set of basic bi-directional flow parameters used in our work is shown in **Table 1**. The Packet CAPture (PCAP) files offered by the MAWI lab were converted into a bi-directional flow. After creating a bi-directional flow, we managed data cleaning by eliminating potential network attacks, which are the flows with zero millisecond duration and made of one packet because those flows would commonly be classified as a network scanning activity [50].

Then, the traffic structures were modeled as a stationary process, inspired by a work of Barakat et al. [11]. To obtain the descriptive model for background traffic, the model fitting technique was applied to find the proper distribution and its corresponding parameters. The time series of flows were fitted with parametric distributions, namely normal, Poisson, exponential, gamma, nbinomial, geom, uniform, and logistic. The maximum likeli-

Table 1 Parameters of flow records.

No.	Name	Explanations
1.	Dur	Flow record duration
2.	Proto	Flow transaction protocol
3.	Sport	Source port number
4.	Dport	Destination port number
5.	TotPkts	Total transaction packet count
6.	SrcPkts	Source - Destination packet count
7.	TotBytes	Total transaction bytes
8.	SrcBytes	Source - Destination transaction bytes
9.	DstBytes	Destination - Source transaction bytes
10.	TotAppBytes	Total application bytes
11.	SAppBytes	Source - Destination application bytes
12.	DAppBytes	Destination - Source application bytes
13.	Rate	Packet count per second
14.	SrcRate	Source packet count per second
15.	DstRate	Destination packet count per second
16.	Dir	Direction of transaction
17.	State	Transaction state
18.	dMaxPktSz	Maximum packet size for traffic transmitted by destination
19.	sMinPktSz	Minimum packet size for traffic transmitted by source
20.	dMinPktSz	Minimum packet size for traffic transmitted by destination

Table 2 Distribution functions used in modeling.

Theoretical Distribution	Cumulative Distribution Function	Parameters
Normal	$F(x) = \int_{-\infty}^{\infty} xf(x)dx$, where $f(x) = e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ, σ
Poisson	$F(x) = e^{-\alpha} \sum_{i=0}^{k} \frac{\lambda^i}{i!}$	λ
Exponential	$F(x) = 1 - e^{\frac{-x}{\mu}}$	μ
gamma	$F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$	α, β
Geometric	$F(x) = 1 - (1 - p)^{k+1}$	p
nbinomial	$F(x) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1 - p)^{n-i}$	n, p
uniform	$F(x) = \left\{ \begin{array}{cc} 0 & \text{for } x - \mu < -\sigma\sqrt{3} \\ \frac{1}{2}\left(\frac{x-\mu}{\sigma\sqrt{3}} + 1\right) & \text{for } -\sigma\sqrt{3} \leqslant x - \mu < \sigma\sqrt{3} \\ 1 & \text{for } x - \mu \geq \sigma\sqrt{3} \end{array} \right\}$	μ, σ
logistic	$\frac{1}{1+e^{-\frac{x-\mu}{s}}}$	μ, s

hood estimator method was applied to estimate the relevant parameters. We selected the distributions that best fit the traffic features of interests, listed as flow duration, packet count per second, and flow size in byte. The goodness-of-fits was examined using Kolmogorov-Smirnov statistics.

In this paper, a background traffic from source s to destination d is characterized by the 3-tuple $\langle D_i, P_i, B_i \rangle$, where D_i is the flow duration. P_i is the packets count per second in a flow. B_i is the flow size in bytes. The distribution functions used in our work are demonstrated in **Table 2**.

Given k univariate distribution $D_1, \ldots, D_k$ and fitness proportionate selection which somehow simulate iid data from each distribution. We wanted to generate a random matrix $B_{nk} = [b_1 \ \ldots \ b_k]$ of each traffic features. By comparing the discrepancy between each of theoretical distribution and its corresponding empirical data, we received the theoretical distribution and the corresponding parameters, needed in regenerating background traffic.

3.2 Attack Traffic Configuration

To create a dataset for IDS evaluation, some attack traffics are necessitated. By adopting the idea of attack profiling proposed by Futoransky et al. [21], we formulated a predefined fitness function for each attack of interest to create the transition of the simulation framework. The attack traffic configurations were also used in crafting network attacks of interest, listed as follows: network scanning, host scanning, web vulnerability scanning, SSH bruteforcing, and SQL injection. This paper focused on generating only 5 important network attacks, but the framework also allows for adding more types of attacks.

3.3 Dataset Generation

The proposed framework applied a discrete event simulation technique, which allowed us to create both benign background traffic and malicious network traffic. The dynamic of the simulation over time was achieved by implementing a fitness proportionate selection algorithm. The fitness function was assigned a fitness to possible solutions and was used to associate a probability of selection with each population [23], [35]. The idea was comparable to a Roulette wheel in which, the proportion of the wheel assigned to each possible selection based on their fitness value. The fitness level was used to associate the probability of selection with each generating event.

If f_i is the fitness of individual i in the population, its probability of being selected will be defined as:

$$p_i = \frac{f_i}{\sum_j^N f_j};$$

where N is the number of individuals in the population. The expected number of an individual type of event generated by pro-

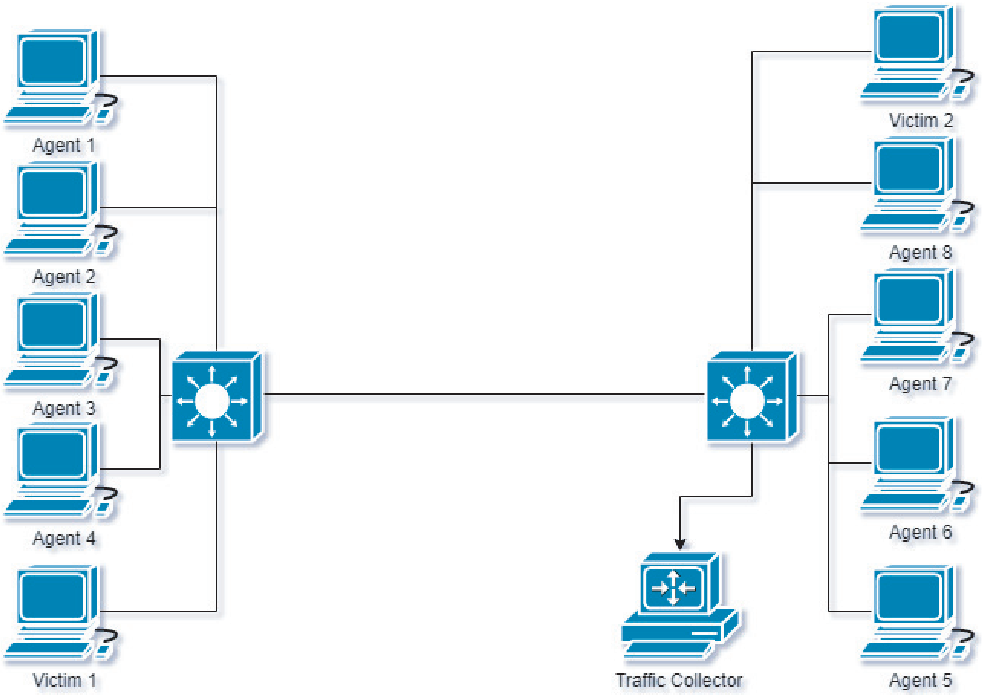

Fig. 3 The experimental network architecture.

Table 3 The MAWI April 9, 2019, 0000-0015 dataset.

2019-04-09 H0000	Total	DNS	FTP	HTTP	SMTP	SSH
Packet	70,676,113	405,699	43,537	16,474,907	244,360	8,684,593
Bi-directional Flow	24,927,010	289,708	8,667	440,722	21,178	121,662

portional selection was defined as:

$$E(N_i)_P = NP_i = \frac{Nf_i}{\sum_{j=1}^{N} f_j}$$

Based on this idea, the expected distribution of event generation and the transition of events for traffic simulation has been defined. The desired distribution functions were expressed as a linear combination of some uniform distribution functions, shown in Eq. (1).

$$F_x(t) = \sum_{i=1}^{n} p_i F_{x_i}(t), \quad \sum_{i=1}^{n} p_i = 1 \tag{1}$$

The discrete simulation prototyping software was developed, as described in Section 3.3. The prototyping software was designed to simultaneously synthesize background and attack traffic. Each event was carried out by generating the cumulative probability distribution (CDF) over the list of events using a probability proportional to the fitness of the event types. The D-ITG [13] was employed to generate the background traffic according to the obtained models and Kali Linux [2] was used as a platform for the attack generations. The network architecture is presented in **Fig. 3**.

4. Experimental Results

The extensive description of the experiments regarding traffic metric extraction, dataset generation, and the exploratory data analysis of the simulated dataset will be presented in the following subsections.

4.1 Background Traffic Models

4.1.1 The Analyzed MAWI Dataset

We analyzed datasets selected from 48-hour-long traces provided by WIDE backbone [31] to portray the Internet background traffic. Each network trace was 15 minutes long, which was collected at the transit link of the WIDE backbone to the upstream ISP. In this paper, network traces captured between April 9, 2019, 0000-0015 is presented. **Table 3** shows the distribution of the selected network trace, summarized by application protocols. The number of packets was provided by the MAWI, while the number of bi-directional flows was obtained from the extraction process, described in Section 3.1. The table shows the different quantity of created flows versus the number of packets concerning each application protocol.

4.1.2 Structural Session Traffic

We concentrated on analyzing flow size, flow duration, and packets count per second to recreate the benign background traffic. To obtain simplistic and non-periodic traffic, probabilistic data propagation was used in a condition that the host generates flow traffic at each specific trigger was based on a given probability. The application traffic structure was determined by the probability value assigned to the focused application properties. **Table 4** shows selected theoretical distributions, their corresponding parameters and the Kolmogorov-Smirnov (KS) score for 201904090000 trace. This proved the statistical difference of bi-directional flow features toward the different application protocols. The fitting results suggested that the flow size, flow duration, and packets count per second tended to have different distributions and the FTP traffic's features could be defined using the

Table 4 Selected Distributions and Relevant Parameters for 201904090000 trace.

	DNS	FTP	HTTP	SMTP	SSH
Packet Count per Second	gamma 0.6866, 0.0175 KS. = 0.28556	unif −1.556, 14.023 KS. = 0.13307	gamma 0.4832, 0.0323 KS. = 0.16288	logistic 3.3366, 1.7399 KS. = 0.12811	nbinom 0.7362, 7.3500 KS. = 0.20480
Flow Duration	gamma 0.6039, 11.6684 KS. = 0.30576	normal 2.1712, 1.6279 KS. = 0.10254	exp 0.568036 KS. = 0.16049	exp 0.50331 KS. = 0.10590	normal 1.7623, 1.6232 KS. = 0.1849
Flow Byte	normal 300.165, 116.294 KS. = 0.09340	normal 1,161.73, 1,129.55 KS. = 0.17265	poisson 3,655.09 KS. = 0.17012	normal 355.830, 307.452 KS. = 0.12362	unif −462.066, 1,866.273 KS. = 0.19845

Table 5 Summary of the generated dataset.

Benign	**Attack**				
	Host Scan	Network Scan	SQL Injection	SSH BruteForce	Web Vulnerability Scan
218,729	7,015	8,055	7,676	7,463	7,632

normal distribution. The experiments also revealed that HTTP and SMTP flow duration seemed alike since they both could be described by the exponential distribution, having a roughly comparable parameter value.

4.2 The Generated Dataset

We validated our framework by matching the synthesized dataset with the MAWI dataset 201904090000. The summary of the generated dataset is shown in **Table 5**. The simulation framework was configured to generate 15 percent of malicious traffic, and yielded a dataset containing 256,633 bidirectional traffic flows, consisting of 218,729 benign bi-directional flows and 37,904 attacks flow. These following attacks were included in the generated dataset, host scanning, network scanning, SQL injection, SSH brute-forcing, and web vulnerability scanning. A dataset simulated from the prototyping simulation could produce a collection of benign and malicious traffic combinations, which demonstrated a satisfying stochastic behavior synthesized by fitness proportionate selection.

4.2.1 Statistical Similarity of the Generated Dataset

The differences between two empirical cumulative distribution functions (ECDF) were measured in order to measure the similarity of the generated dataset and the MAWI dataset. The ECDF plots of traffic statistics derived from 201904090000 trace, depicted in **Figs. 4**, **5** and **6**. The MAWI ECDF was used as a reference distribution to measure the traffic feature, simulated by the framework. The figures confirm that the structures of simulated HTTP traffic closely agree with the MAWI201904090000 bi-directional flow. Therefore, it has been shown that our method can simulate realistic and dynamic traffic features of bi-directional background traffic flow.

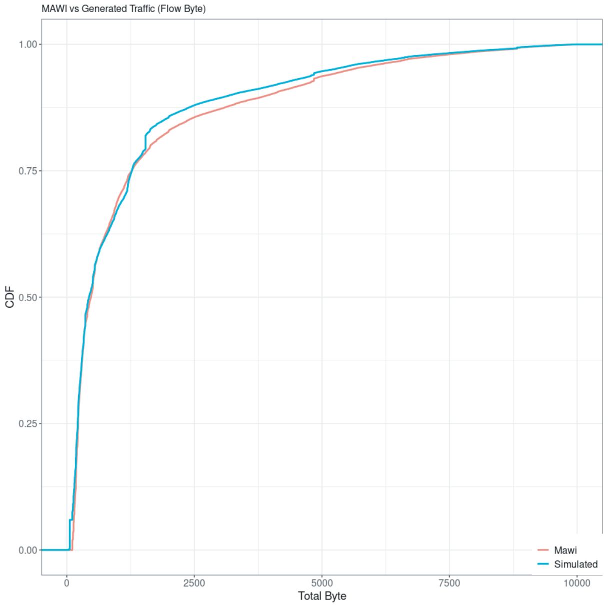

Fig. 4 The ECDF of modeled HTTP flow byte VS simulated. It is clear that the empirical cumulative distribution function of generated HTTP flow byte and MAWI is similar.

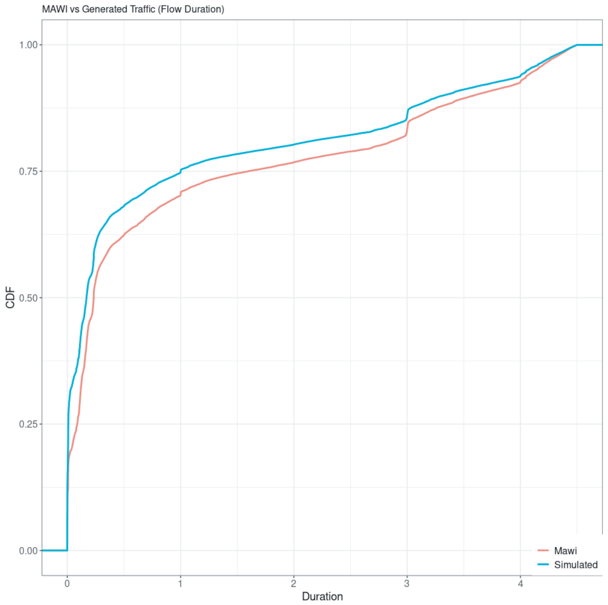

Fig. 5 The ECDF of modeled HTTP flow duration VS simulated. It is obvious that the empirical cumulative distribution function of generated HTTP flow duration and MAWI is comparable.

4.2.2 Exploratory Data Analysis via PCA

In this section, our exploratory data analysis on the simulated dataset was summarized using its main characteristics and the relationships between the bi-directional flow traffic features were visualized. The principal component analysis (PCA) was employed to investigate the linear correlation between the bi-directional flow traffic feature, persisting in our simulated dataset.

The relationship between each feature in a dataset could be shown using the variable factor map [33]. The factor map presents the amount of variance from each traffic feature on the

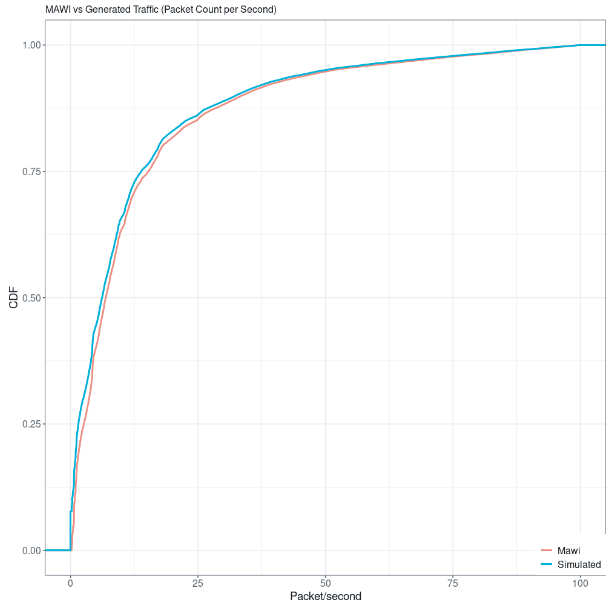

Fig. 6 The ECDF of modeled HTTP packet count/sec VS simulated. It is apparent that the empirical cumulative distribution function of generated HTTP packet count per second and MAWI is similar.

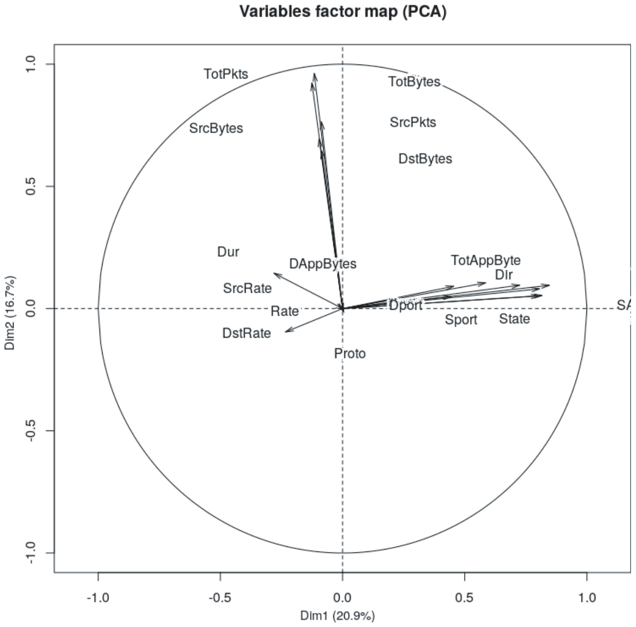

Fig. 7 The variable factor map. TotPkts, SrcBytes, DAppBytes, Dur, SrcByte, Rate, DstRate, and Proto showed a negative correlation in the first dimension.

total variance in the principal component, so we performed principal component analysis on the dataset and visualized its factor map, as shown in **Fig. 7**. The first dimension explained approximately 20% of the total variation considering all traffic features. TotAppByte, Dir, Dport, Sport, and State were described as strong positively correlated. However, TotPkts, SrcBytes, DAppBytes, Dur, SrcByte, Rate, DstRate, and Proto showed a negative correlation in the first dimension.

The second dimension explained approximately 17% of the total variance and revealed that some parameters exhibited a different influence on the total variation. In this dimension, TotPkts and SrcBytes show a robust positive relationship, while DstRate and Rate expressed the low negative impact on the total variation.

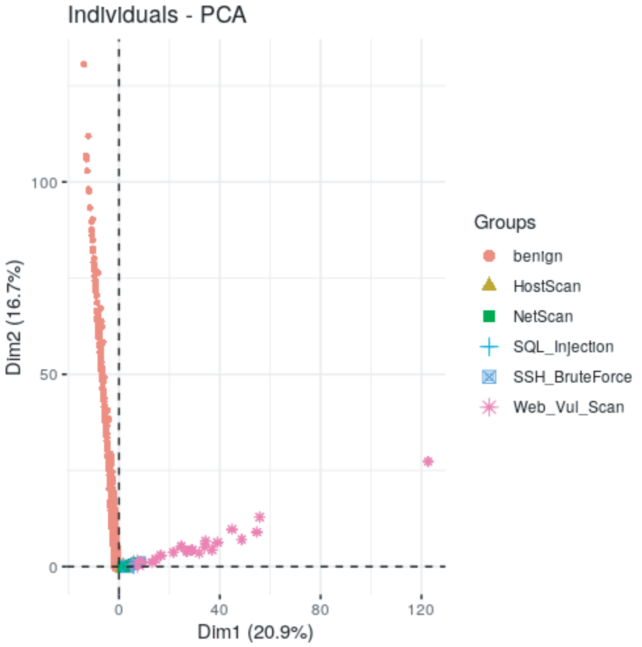

Fig. 8 The individual plot of the first two dimensions. The cluster of benign traffic separates from malicous traffics.

The individual plot, shown in **Fig. 8** also showed that the benign traffic appeared to have the weak negative correlation of TotPkts, SrcBytes, DAppBytes, Rate in the first principal component, while the strong correlation of TotAppByte, Dir, Dport, Sport, and State revealed to signify the characteristic of web vulnerability scanning, which explicated that the characteristic of benign web traffic was different from the web vulnerability scanning activities, even if they were frequently seen as web traffic. In addition, the plot also revealed the clusters of network attacks, showing that the generated traffic was suitable for IDS evaluation.

5. Conclusion

We believe that the proposed framework will be useful to researchers who work on bi-directional traffic simulation or network security research community. The framework can generate a good dataset for bi-directional flow-based IDS evaluation. It is proficient to replicate realistic background traffic from real-world network measurements using our suggested structural bi-directional flow traffic models. The main improvement over others is that the background traffic is very representative of various application protocol traffics, which hold a comparable characteristics, reflecting DNS, FTP, HTTP, SMTP, and SSH traffics. Furthermore, the malicious traffic generation utilizes the fitness proportionate selection, where each attack of interest is mapped to the probability of selection, providing a better stochastic means to generate malicious traffics. This allows other attacks of interest to be easily configured and used to regenerate the dataset with different properties, allowing the insertion of more complicated network attacks to create a dataset that copes with a brand-new attack and shares with other researchers.

There are different efforts to generate realistic flow-based datasets, which are regularly developed based on the unidirectional flow modeling with network features such as flow duration, the flow size, and packets are usually modeled. In this paper, we especially consider generating realistic bi-directional background centering on structural session traffic features, as well as implanting the up-to-date attacks. We believe that this is the first time that one of the bi-directional flow-based features – Rate (packet-count per second), has been applied to generate realistic bi-directional background traffic. Supported by the exploratory data analysis results, there is the significant importance of traffic rate, persisting in benign web traffic that suggests additional research efforts to evaluate its impact toward detecting attacks against web services. The experimental result illustrates that the characteristics of benign web traffics are different from the vulnerability scanning activities.

In this paper, the attack and bi-directional traffic are systematically organized, recognized, and achieved by our approach for generating a dataset for bi-directional flow-based IDS evaluation. Some gaps need to be filled, for example, embracing attacks at various kinds of network traffics and striving against advanced network infrastructures. We leave for the future additional studies such as intelligent attacker modeling and distributed simulation.

References

[1] RFC5103: Bidirectional Flow Export Using IP Flow Information Export (online), available from ⟨https://tools.ietf.org/html/rfc5103⟩ (accessed 2020-03-09).

[2] Kali: Our Most Advanced Penetration Testing Distribution, Ever. (online), available from ⟨https://www.kali.org/⟩ (accessed 2020-03-09).

[3] Snort: Snort – An open source network intrusion prevention and detection system (online), available from ⟨https://www.snort.org/⟩ (accessed 2020-03-09).

[4] iCTF: The UCSB The International Capture The Flag (online), available from ⟨https://ictf.cs.ucsb.edu/⟩ (accessed 2020-03-09).

[5] KDD99: KDD Cup 1999 Data (online), available from ⟨http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html⟩ (accessed 2020-03-09).

[6] Suricata: Open Source IDS / IPS / NSM engine (online), available from ⟨http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html⟩ (accessed 2020-03-09).

[7] Afanasiev, F., Daly, W., Sukhov, A. and Petrov, A.: Flow-based analysis of Internet traffic, *Computing Research Repository*, Vol.cs.NI/0306, pp.92–95 (2003).

[8] Aikat, J., Hasan, S., Jeffay, K. and Smith, F.D.: Towards Traffic Benchmarks for Empirical Networking Research: The Role of Connection Structure in Traffic Workload Modeling, *2012 IEEE 20th International Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems* (*MASCOTS*), pp.78–86, IEEE (2012).

[9] Antonello, R., Fernandes, S., Kamienski, C., Sadok, D., Kelner, J., Gódor, I., Szabó, G. and Westholm, T.: Deep packet inspection tools and techniques in commodity platforms: Challenges and trends, *Journal of Network and Computer Applications*, Vol.35, No.6, pp.1863–1878 (2012).

[10] Antonenko, V.A. and Smelyanskiy, R.L.: Simulation of Malicious Activity in Wide Area Networks, *Programming and Computer Software*, Vol.39, No.1, pp.25–33 (2013).

[11] Barakat, C., Thiran, P., Iannaccone, G., Diot, C. and Owezarski, P.: Modeling Internet backbone traffic at the flow level, *IEEE Trans. Signal Processing*, Vol.51, No.8, pp.2111–2124 (2003).

[12] Boschi, E. and Trammell, B.: Bidirectional Flow Measurement, IPFIX, and Security Analysis, *FloCon 2006*, Portland, OR, CERT (2006).

[13] Botta, A., Dainotti, A. and Pescapé, A.: A tool for the generation of realistic network workload for emerging networking scenarios, *Computer Networks*, Vol.56, No.15, pp.3531–3547 (2012).

[14] Bye, R., Schmidt, S., Luther, K. and Albayrak, S.: Application-level simulation for network security, *Proc. 1st International ICST Conference on Simulation Tools and Techniques for Communications Networks and Systems* (2008).

[15] Chen, T.M.: Network Traffic Modeling, *Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, 3rd edition, pp.326–339, John Wiley & Sons, Inc. (2007).

[16] Dowoo, B., Jung, Y. and Choi, C.: PcapGAN: Packet capture file generator by style-based generative adversarial networks, *Proc. 18th IEEE International Conference on Machine Learning and Applications*, *ICMLA 2019*, pp.1149–1154 (2019).

[17] Elejla, O.E., Anbar, M., Belaton, B. and Hamouda, S.: Labeled flow-based dataset of ICMPv6-based DDoS attacks, *Neural Computing and Applications* (2018).

[18] Erramilli, V., Crovella, M. and Taft, N.: An independent-connection model for traffic matrices, *Proc. 6th ACM SIGCOMM on Internet Measurement – IMC '06*, p.215,, ACM Press (2006).

[19] Floyd, S. and Paxson, V.: Difficulties in simulating the Internet, *IEEE/ACM Trans. Networking*, Vol.9, No.4, pp.392–403 (2001).

[20] Fontugne, R., Borgnat, P., Abry, P. and Fukuda, K.: MAWILab, *Proc. 6th International Conference on – Co-NEXT '10*, ACM Press (2010).

[21] Futoransky, A., Miranda, F., Orlicki, J. and Sarraute, C.: Simulating cyber-attacks for fun and profit, *Proc. 2nd International ICST Conference on Simulation Tools and Techniques*, ICST (2009).

[22] Gao, M., Zhang, K. and Lu, J.: Efficient packet matching for gigabit network intrusion detection using TCAMs, *20th International Conference on Advanced Information Networking and Applications – Volume 1* (*AINA '06*), p.6, IEEE (2006).

[23] Gen, M., Cheng, R. and Lin, L.: *Network Models and Optimization*, Decision Engineering, Springer London (2008).

[24] Gogoi, P., Bhuyan, M., Bhattacharyya, D. and Kalita, J.: Packet and Flow Based Network Intrusion Dataset, *Contemporary Computing*, Parashar, M., Kaushik, D., Rana, O.F., Samtaney, R., Yang, Y. and Zomaya, A. (Eds.), pp.322–334, Springer Berlin Heidelberg (2012).

[25] Golaup, A. and Aghvami, H.: A multimedia traffic modeling framework for simulation-based performance evaluation studies, *Computer Networks*, Vol.50, No.12, pp.2071–2087 (2006).

[26] Gomes, J.V.P., Inácio, P.R.M., Lakic, B., Freire, M.M., Da Silva, H.J.A. and Monteiro, P.P.: Source traffic analysis, *ACM Trans. Multimedia Computing, Communications, and Applications*, Vol.6, No.3, pp.1–23 (2010).

[27] Guo, C., Zhou, Y.-J., Ping, Y., Luo, S.-S., Lai, Y.-P. and Zhang, Z.-K.: Efficient intrusion detection using representative instances, *Computers & Security*, Vol.39, pp.255–267 (2013).

[28] Haider, W., Hu, J., Slay, J., Turnbull, B.P. and Xie, Y.: Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling, *Journal of Network and Computer Applications*, Vol.87, No.March, pp.185–192 (2017).

[29] Hoflack, L., Vuyst, S.D., Wittevrongel, S. and Bruneel, H.: Modeling Web Server Traffic with Session-Based Arrival Streams, *Analytical and Stochastic Modeling Techniques and Applications*, Al-Begain, K., Heindl, A. and Telek, M. (Eds.), Lecture Notes in Computer Science, Vol.5055, pp.47–60, Springer Berlin Heidelberg (2008).

[30] Issariyakul, T. and Hossain, E.: *Introduction to Network Simulator NS2*, 1st edition, Springer Publishing Company, Inc. (2008).

[31] Cho, K., Mitsuya, K. and Kato, A.: Traffic Data Repository at the WIDE Project, *FREENIX Track*, pp.263–270, USENIX Association (2000).

[32] Kim, M.-S., Won, Y.J. and Hong, J.W.: Characteristic analysis of internet traffic from the perspective of flows, *Computer Communications*, Vol.29, No.10, pp.1639–1652 (2006).

[33] Lê, S., Josse, J. and Husson, F.: FactoMineR : An R Package for Multivariate Analysis, *Journal of Statistical Software*, Vol.25, No.1, pp.1–18 (2008).

[34] Lee, I.W. and Fapojuwo, A.O.: Stochastic processes for computer network traffic modeling, *Computer Communications*, Vol.29, No.1, pp.1–23 (2005).

[35] Lipowski, A. and Lipowska, D.: Roulette-wheel selection via stochastic acceptance, *Physica A: Statistical Mechanics and Its Applications*, Vol.391, No.6, pp.2193–2196 (2012).

[36] Lippmann, R., Fried, D.J., Graf, I., Haines, J.W., Kendall, K.R., Mcclung, D., Weber, D., Webster, S.E., Wyschogrod, D., Cunningham, R.K., Zissman, M.A. and Street, W.: Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation, *the 2000 DARPA Information Survivability Conference and Exposition* (*DISCEX*), IEEE Computer Society (2000).

[37] Lippmann, R., Haines, J.W., Fried, D.J., Korba, J. and Das, K.: Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation, *Computer Networks*, Vol.34, pp.162–182, Springer Verlag (2000).

[38] Maciá-Fernández, G., Camacho, J., Magán-Carrión, R., García-Teodoro, P. and Therón, R.: UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs, *Computers and Security*, Vol.73, pp.411–424 (2018).

[39] Paxson, V.: Bro: A system for detecting network intruders in real-time, *Computer Networks*, Vol.31, No.23-24, pp.2435–2463 (1999).

[40] Ring, M., Schlör, D., Landes, D. and Hotho, A.: Flow-based network traffic generation using Generative Adversarial Networks, *Computers and Security*, Vol.82, pp.156–172 (2019).

[41] Roughan, M., Lund, C. and Donoho, D.: Estimating point-to-point and point-to-multipoint traffic matrices: An information-theoretic approach, *IEEE/ACM Trans. Networking*, Vol.13, No.5, pp.947–960 (2005).

[42] Sarraute, C., Miranda, F. and Orlicki, J.I.: Simulation of Computer Network Attacks, *CoRR*, Vol.abs/1006.2, p.17 (2010).

[43] Schmidt, S., Bye, R., Chinnow, J., Bsufka, K., Camtepe, A. and Albayrak, S.: Application-level Simulation for Network Security, *Simulation*, Vol.8, No.5-6 (2010).

[44] Sharma, R., Singla, R.K. and Guleria, A.: A New Labeled Flow-based DNS Dataset for Anomaly Detection: PUF Dataset, *Procedia Computer Science*, Vol.132, pp.1458–1466 (2018).

[45] Shiravi, A., Shiravi, H., Tavallaee, M. and Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection, *Computers & Security*, Vol.31, No.3, pp.357–374 (2012).

[46] Sommers, J., Bowden, R., Eriksson, B., Barford, P., Roughan, M. and Duffield, N.: Efficient network-wide flow record generation, *2011 Proc. IEEE INFOCOM*, pp.2363–2371, IEEE (2011).

[47] Song, L. and Gerald, M.A.: Realistic Internet Traffic Simulation Through Mixture Modeling and a Case Study, *Proc. Winter Simulation Conference, 2005*, pp.2408–2416, IEEE (2005).

[48] Song, Y., Stolfo, S.J. and Jebara, T.: Behavior-based network traffic synthesis, *2011 IEEE International Conference on Technologies for Homeland Security*, *HST 2011*, pp.338–344 (2011).

[49] Soule, A., Nucci, A., Cruz, R., Leonardi, E. and Taft, N.: How to identify and estimate the largest traffic matrix elements in a dynamic environment, *Proc. Joint International Conference on Measurement and Modeling of Computer Systems - SIGMETRICS 2004/PERFORMANCE 2004*, p.73, ACM Press (2004).

[50] Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C., Pras, A. and Stiller, B.: An Overview of IP Flow-Based Intrusion Detection, *IEEE Communications Surveys & Tutorials*, Vol.12, No.3, pp.343–356 (2010).

[51] Umer, M.F., Sher, M. and Bi, Y.: Flow-based intrusion detection: Techniques and challenges, *Computers & Security*, Vol.70, pp.238–254 (2017).

[52] Vasudevan, A., Harshini, E. and Selvakumar, S.: SSENet-2011: A Network Intrusion Detection System dataset and its comparison with KDD CUP 99 dataset, *2011 2nd Asian Himalayas International Conference on Internet* (*AH-ICI*), pp.1–5, IEEE (2011).

[53] Viegas, E.K., Santin, A.O. and Oliveira, L.S.: Toward a reliable anomaly-based intrusion detection in real-world environments, *Computer Networks*, Vol.127, pp.200–216 (2017).

[54] Vishwanath, K. and Vahdat, A.: Swing: Realistic and Responsive Network Traffic Generation, *IEEE/ACM Trans. Networking*, Vol.17, No.3, pp.712–725 (2009).

[55] Wang, H., Gong, Z., Guan, Q. and Wang, B.: Detection Network Anomalies Based on Packet and Flow Analysis, *7th International Conference on Networking* (*ICN 2008*), pp.497–502 (2008).

[56] Wette, P. and Karl, H.: DCT 2 Gen: A traffic generator for data centers, *Computer Communications*, Vol.80, pp.45–58 (online), DOI: 10.1016/j.comcom.2015.12.001 (2016).

[57] Yang, W., Gong, J., Ding, W. and Wu, X.: Network Traffic Emulation for IDS Evaluation, *2007 IFIP International Conference on Network and Parallel Computing Workshops* (*NPC 2007*), pp.608–612, IEEE (2007).

Korakoch Wilailux is serving in Royal Thai Navy as a military instructor in Division of Academic Affairs, Department of Naval Education. He received his B.Sc. degree from Royal Thai Naval Academy in 2000 and his M.Sc. from Mahidol University in 2006. He is currently a Ph.D. Candidate in the Faculty of ICT, Mahidol University. His research focuses on network security, information warfare, and Cyberwar.

Sudsanguan Ngamsuriyaroj is an Associate Professor in the Faculty of ICT, Mahidol University, Thailand. She received her Ph.D. in Computer Science and Engineering from Pennsylvania State University in 2002. Her research interests are network security, high-performance computing, and performance evaluation.