Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling☆

W. Haider[a], J. Hu[a,*,1], J. Slay[a], B.P. Turnbull[a], Y. Xie[b]

[a] *School of Engineering and Information Technology, University of New South Wales at Australian Defence Force Academy, Canberra, Australia*
[b] *School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, PR China*

ARTICLE INFO

Keywords:
IDS
IDS dataset
Dataset evaluation
Dataset realism
Fuzzy logic
HIDS
NIDS

ABSTRACT

Prior to deploying any intrusion detection system, it is essential to obtain a realistic evaluation of its performance. However, the major problems currently faced by the research community is the lack of availability of any realistic evaluation dataset and systematic metric for assessing the quantified quality of realism of any intrusion detection system dataset. It is difficult to access and collect data from real-world enterprise networks due to business continuity and integrity issues. In response to this, in this paper, firstly, a metric using a fuzzy logic system based on the Sugeno fuzzy inference model for evaluating the quality of the realism of existing intrusion detection system datasets is proposed. Secondly, based on the proposed metric results, a synthetically realistic next generation intrusion detection systems dataset is designed and generated, and a preliminary analysis conducted to assist in the design of future intrusion detection systems. This generated dataset consists of both normal and abnormal reflections of current network activities occurring at critical cyber infrastructure levels in various enterprises. Finally, using the proposed metric, the generated dataset is analyzed to assess the quality of its realism, with its comparison with publicly available intrusion detection system datasets for verifying its superiority.

1. Introduction

Since the beginning of the 1970s, intrusion detection systems (IDSs) have been extensively adopted to protect computer networks against both known and unknown attacks (Ahmed et al., 2016; Hu et al., 2011; Haider et al., 2015). A realistic audit dataset of computer networks plays the role of evaluating the performance (e.g. accuracy and computational time) of an IDS design (Hoang et al., 2009; Haider et al., 2016). DARPA KDD98, which was generated approximately eighteen years ago and is considered the gold standard of datasets, has become outdated in terms of real-world networks normal traffic and attack behaviors (Haider et al., 2015; Creech, 2014). Although several related reports (Zuech et al., 2015; Shiravi et al., 2012; Gogoi et al., 2012; McHugh, 2000; Vasudevan et al., 2011; Sangster et al., 2009; Song et al., 2011) have attempted to extend and replace this dataset, but they have been unsuccessful because: (i) due to concerns regarding their integrity, cost and interruption of business operations, enterprises have not allowed IDS researchers to use their actual production networks for the purpose of penetration testing; and (ii) it is practically impossible for the low-resource-enabled testbeds or home-made honeypots used by existing IDS dataset generators to simulate normal traffic dynamics and possible attacks of real-world networks. Moreover, the quantified quality of realism of any IDS dataset cannot be assessed using the metric given in Shiravi et al. (2012) which is based on YES/NO arguments regarding the incompletely defined design features of an IDS dataset generation process.

To fill these gaps, in Section 2, we propose a metric based on a fuzzy logic system (FLS) that provides a theory for evaluating the quality of realism of any IDS dataset. Then, it is used to evaluate existing IDS datasets to demonstrate its capability to quantify their quality of realism. In Section 3, current IDS datasets and their quality evaluation criteria are briefly discussed to justify the need for the proposed metric and IDS dataset. Based on the results for the quality of realism evaluation obtained by the proposed metric, in Section 4, we briefly describe how the authors designed and generated a synthetically realistic IDS dataset, i.e., the next-generation IDS dataset (NGIDS-DS) using IXIA *Perfect Storm* (Penetration Tester, 2016) in conjunction with a range of commercial cyber-security-test hardware platforms. Also, the quantity of captured data and preliminary analysis of the simulation traffic are provided. In Section 5, firstly, we apply the proposed metric on the NGIDS-DS to calculate its quality of realism and compare it with those of existing datasets which proves that it is superior.

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.

E-mail address: J.hu@adfa.edu.au (J. Hu).

[1] IEEE Senior member.

http://dx.doi.org/10.1016/j.jnca.2017.03.018
Received 2 November 2016; Received in revised form 6 March 2017; Accepted 20 March 2017
Available online 28 March 2017

Secondly, the IDS dataset quality evaluation metric in Shiravi et al. (2012) is also applied on the proposed and existing IDS datasets to illustrate the difference between it and the proposed metric.

2. Evaluation of realism of IDS datasets

The need for the NGIDS-DS is justified through an IDS dataset quality of realism evaluation metric modeled using a FLS based on the Sugeno fuzzy inference engine (Sugeno and Yasukawa, 1993) which consists of four main parts: a fuzzifier; rules; an inference engine; and a defuzzifier. From the perspective of qualitative modeling, it is considered that fuzzification is the process of quantifying a qualitative subject, in this case, the quality of realism of an IDS dataset, by defining its components, i.e., an inference engine, crisp sets of input data, rules, membership functions, fuzzy linguistic variables, a fuzzy set and fuzzy linguistic terms.

The Sugeno fuzzy inference model, which is adopted as the FLS's inference engine, requires two distinct crisp input sets and the relationship(s) between their elements to form rules. Accordingly, we define these sets as $X = \{x_1, x_2, x_3 \ldots x_6\}$ and $Y = \{y_1, y_2\}$, and their input membership functions as $F_1(x_k)$ and $F_2(y_l)$ in Tables 1, 2 respectively. Set X represents the factors of a possible realistic IDS dataset acquired in chunks from related efforts (Creech and Hu, 2013; Zuech et al., 2015; Shiravi et al., 2012; Gogoi et al., 2012; McHugh, 2000; Vasudevan et al., 2011; Sangster et al., 2009; Song et al., 2011) and set Y the environmental variable for generating an IDS dataset (Davis and Magrath, 2013). Importantly, in Table 1, for factor x_4, the operational timing refers to the inclusion in the IDS dataset of peak hours, after hours, nights, weekends and time zone differences, and the industry complexity refers the varieties of cyber-based enterprises, such as communications, the military, banks, academia, health, social media, e-commerce as well as number of users and applications. Further, there are two reasons to propose and define the elements of set X and Y: (i) to establish the minimal ingredients to generate an IDS dataset; and (ii) to select and evaluate the quality of realism of any IDS dataset before applying to an IDS design.

The task of the membership function $F_1(x_k)$ is to assign a predefined output a singleton value, i.e., 0.16, where the value of k is dependent on the particular dataset under observation and the factor(s) from set X in it. It is determined by the fact that, if a dataset has maximum realism, the probability of realism is 1. Furthermore, if we express the maximum realism as 1 with respect to the contributions of the six predefined factors, for each member of set X, its share in contributing to realism will be 1/6. On the other hand, the purpose of the membership function $F_2(y_l)$ is to assign a predefined output a singleton value, i.e., 1 for real or 0.5 for synthetic to input y_l, where l is dependent on the particular dataset under observation and its type of generation environment. It actually embeds an approximation of the fact into the rule inference, i.e., the probability of the quality of realism of an IDS dataset generated over real networks will be considered the maximum and half over the synthetic ones. Moreover, the predefined assignment values (e.g., 0.16, 1 and 0.5) of the membership functions (i.e., $F_1(x_k)$ and $F_2(y_l)$) can be re-enumerated after extending or defining the elements of set X and Y.

Table 1
Crisp input set X.

Elements	Description	$F_1(x_k)$
x_1	Complete capture of audit logs of computer operating system and network packets	0.16
x_2	Maximum number of possible attacks included	0.16
x_3	Current attack behaviors	0.16
x_4	Real-world normal traffic dynamics with operation timings and industry complexity	0.16
x_5	Maintenance of cyber infrastructure performance during complete capture	0.16
x_6	Ground truth information included to assist labeling process	0.16

Table 2
Crisp input set Y.

Elements	Linguistic terms	Generation environment	$F_2(y_l)$
y_1	Good	Production or real network	1
y_2	Average	Synthetic network or testbed	0.5

$$z_i = a[F_1(x_k)] + b[F_2(y_l)] + c \tag{1}$$

$$w_i = AndMethod\ [F_1(x_k), F_2(y_l)] \tag{2}$$

In the Sugeno fuzzy inference model, the output level of rule i (z_i) is weighted by its ring strength (w_i), as defined in Eqs. (1), (2) and (3) respectively. As the scaling output parameters (a, b and c in Eq. (1)) for the output level (z_i) are observed to be directly proportional to the number of rules (N), $a = b = c = N$. Then, the final output from the Sugeno model is the weighted average of all the rule outputs, that is,

$$Final\ Output = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \tag{3}$$

The final output is considered the initial numerical value of the realism obtained, i.e., numerical R. In order to achieve the final quantification of the R of an IDS dataset in fuzzy linguistic terms, firstly, the numerical R is normalized as $0 \le R \le 1$. Secondly, the R is considered a fuzzy linguistic variable and equals $\{\text{no, low, medium, medium high, high}\}$ as a fuzzy set, with each member covering a portion of the overall values of the realism as a fuzzy linguistic term. Finally, to map the final R using fuzzy linguistic terms, the five members of fuzzy set R are related to the overlapping ranges of the probability of the realism respectively, i.e., $0 \le R < 0.10 \Longrightarrow \text{no}$, $0.30 \ge R > 0.08 \Longrightarrow \text{low}$, .

$0.60 \ge R > 0.28 \Longrightarrow \text{medium}$,

$0.97 \ge R > 0.58 \Longrightarrow \text{medium high}$ and $1 \ge R > 0.95 \Longrightarrow \text{high}$

The final values of the metric obtained from processing current datasets are shown in Table 3. Each dataset is first analyzed to observe the rules using the finding elements in the defined sets X and Y and their relationships. Based on the observed rules and their quantities N, further calculations are performed using Eqs. (1) and (2) with $i = 1 \ldots N$ to form the numerical R. The final R is obtained through normalization by dividing the numerical R by the maximum numerical R which is calculated as 12.96 using Eqs. (1), (2) and (3) respectively and realizing all the effective rules, i.e., $\{x_1 \text{ AND } y_1, x_2 \text{ AND } y_1, \ldots, x_6 \text{ AND } y_1\}$. In Table 3, the final R values show the probabilities of the quality of realism of existing IDS datasets. Then, to complete the fuzzification step, these values are mapped according to the fuzzy linguistic terms of fuzzy set R and plotted in Fig. 6, where each existing IDS dataset is shown linguistically as having a low or medium quality of realism.

To explain how the above metric works, we provide the following example. Let the final output from Eq. (3) be a network's performance quality, input X its management service and input Y its hardware cost. In fuzzy linguistic terms, the management service can be good, excellent or poor, the hardware cost cheap or expensive and the performance quality high, average or low. In terms of numerical values, good=8, excellent=10, poor=0, expensive=8 and cheap=2. Let us consider a rule for understanding the workings of Eqs. (1), (2) and (3) respectively, e.g., if a network's management service is poor or hardware cost cheap, its performance quality will be low. Let z and w be the numerical inference of this rule and the firing strength respectively. In order to numerically calculate z_i and w_i for the given rule using Eqs. (1) and (2), the input membership functions $F_1(x_k)$ and $F_2(y_l)$ assign the output singleton value as 0 for input X (i.e., the network's management service is poor) and 2 for input Y (i.e., the hardware is cheap). Accordingly, the final output from Eq. (3) will be 0

Table 3
Proposed metric to quantify quality of realism in existing IDS datasets.

Datasets	N	Observed rules	z_i	w_i	numerical R	final R
DARPA McHugh (2000)	2	(x_1 AND y_1)	z_1=4.32	w_1=0.16	4.32	0.33
		(x_6 AND y_1)	z_2=4.32	w_2=0.16		
ISCX Shiravi et al. (2012)	3	(x_1 AND y_1)	z_1=6.48	w_1=0.16	6.18	0.47
		(x_3 AND y_2)	z_2=4.98	w_2=0.08		
		(x_6 AND y_1)	z_3=6.48	w_3=0.16		
ADFA-LD Creech and Hu (2013)	2	(x_3 AND y_2)	z_1=3.32	w_1=0.08	3.98	0.30
		(x_6 AND y_1)	z_2=4.32	w_2=0.16		
Kyoto Song et al. (2011)	1	(x_6 AND y_1)	z_1=2.16	w_1=0.16	2.16	0.16
DEFCON Shiravi et al. (2012)	1	(x_1 AND y_1)	z_1=2.16	w_1=0.16	2.16	0.16
LBNL and CAIDA Shiravi et al. (2012)	1	(x_4 AND y_1)	z_1=2.16	w_1=0.16	2.16	0.16

which can be linguistically modeled as a poor-quality network performance.

We provide three justifications for using a FLS based on the Sugeno fuzzy inference engine while avoiding simple statistical scoring methods. Firstly, quantifying the quality of realism of an IDS dataset is considered a qualitative subject modeling problem. Secondly, fortunately, the Sugeno fuzzy inference model offers a mechanism for calculating the numerical value of R using the problem's components, such as crisp input sets (X and Y), rules and membership functions. Thirdly, the process for estimating the overall score for the degree of realism of a dataset cannot be represented as a single weighted-sum equation due to the dynamics between its input factors (i.e., the elements of X and Y). In other words, the overall rating can be significantly affected by a combination of the subsets of the input elements, for example, we can say that, if the dataset is generated over a real network and contains the maximum number of characteristics of set X, it has a high realism score. Therefore, it is more flexible to represent the scoring process as a combination of rules that judges the subsets of the input sets.

3. Related work

In this section, some current popular IDS datasets and ways of evaluating their quality are briefly discussed with the purpose of answering the following five research questions in order to justify the need for the proposed metric and IDS dataset. (i) How is any particular IDS dataset generated? (ii) How many factor pairs (e.g., joint elements of the proposed crisp input sets X and Y given in Tables 1, 2 respectively; for example, x_1, y_1 is one pair) are included in the design process of any existing IDS dataset? (iii) In a particular current IDS dataset, what are the reasons for a lack of the required factors and their impact on the quality of realism and reliability of an IDS design? (iv) What means or metric is currently used to assess the quality of an IDS dataset? (v) What are the potential problems associated with specific way of assessing the quality of an IDS dataset?

The first constructive effort to generate an IDS dataset conducted by DARPA in 1998 produced KDD98. DARPA used American air force labs to establish a testbed for collecting a real network's normal data and simulating attack traffic. The design process included the two required factors such as x_1, y_1 and x_6, y_1. These datasets are outdated that leads to the lack of other possible required factors pairs (McHugh, 2000). As there is a remarkable difference between the current and past (1998) cyber world. For instance, internet activities in terms of different operating systems, applications and attack tools. Also, there is no information regarding factor x_5 in any publicly available DARPA documentation related to its IDS datasets.

In 2006, Koyoto University Japan released their IDS dataset with the name Kyoto2006 (Song et al., 2011). This dataset is collected by establishing the handmade diverse testbed of honeypots and normal traffic generation servers. The collection of real attack patterns from a network and synthetic generation of normal traffic were the key design features of Kyoto2006 IDS dataset. However, according to our analysis it include x_6, y_1 as the only required factors pair. Moreover, there are several drawbacks with Kyoto 2006: (i) The attack activities and normal traffic were generated separately in two different and uncorrelated environments, which are observed unrealistic collectively in terms of traffic generation and acquisition of logs; and (ii) It is 11 year old which is also outdated.

DEFCON, LBNL and CAIDA are also the publicly available IDS datasets Shiravi et al. (2012). DEFCON was generated by collecting the normal and abnormal traffic while conducting hacking and anti hacking competitions, which is observed, a collection upon restrictive environment. On the other hand, LBNL and CAIDA IDS datasets are collections of normal and abnormal traffic obtained from the internet backbones of large-scale enterprises. DEFCON include x_1, y_1 whereas LBNL and CAIDA consist of only x_4, y_1. Furthermore, the unavailability of commercial-level penetration testing tools and the constraints of enterprises needing business continuity result in a lack of the other required factors pairs in these datasets. Also, unfortunately both LBNL and CAIDA do not provide proper labeling information publicly which is needed for IDS evaluation.

In 2012, the Information Security Centre of Excellence at the University of New Brunswick released its IDS dataset called ISCX (Shiravi et al., 2012). It was generated over realistic network configurations by simulating a real network's possible normal and abnormal scenarios using a group of humans in generating traffic. Based on its documentation, we determined that its design process include the following required factors pairs: (i) x_1, y_1; (ii) x_3, y_2 and (iii) x_6, y_1. The lack of access to an enterprise-level real production network and commercial-level penetration testing tools refects the unavailability of other possible required factor pairs in the ISCX dataset.

Recently, the Australian Defence Force Academy in University of New South Wale has released the two IDS datasets which are the Australian Defence Force Academy Linux Dataset (ADFA-LD) (Creech and Hu, 2013) and University of New South Wale Network Based 2015 (UNSW-NB15) IDS dataset (Moustafa and Slay, 2015, 2016). ADFA-LD consists of normal and abnormal Linux based system calls traces. This dataset was generated via emulation for the evaluation of host based intrusion detection systems. ADFA-LD design process holds the required factors pairs which include x_3, y_2 and x_6, y_1. Its collection of only host logs, hand-made testbed and lack of access to enable data from a real-world enterprise network to be collected are the main reasons for other required factor pairs to be missing in the ADFA-LD. In contrast, UNSW-NB15 is specifically generated over the commercial scale penetration testing environment for the Network based Intrusion Detection Systems (NIDS) evaluation. This dataset includes the

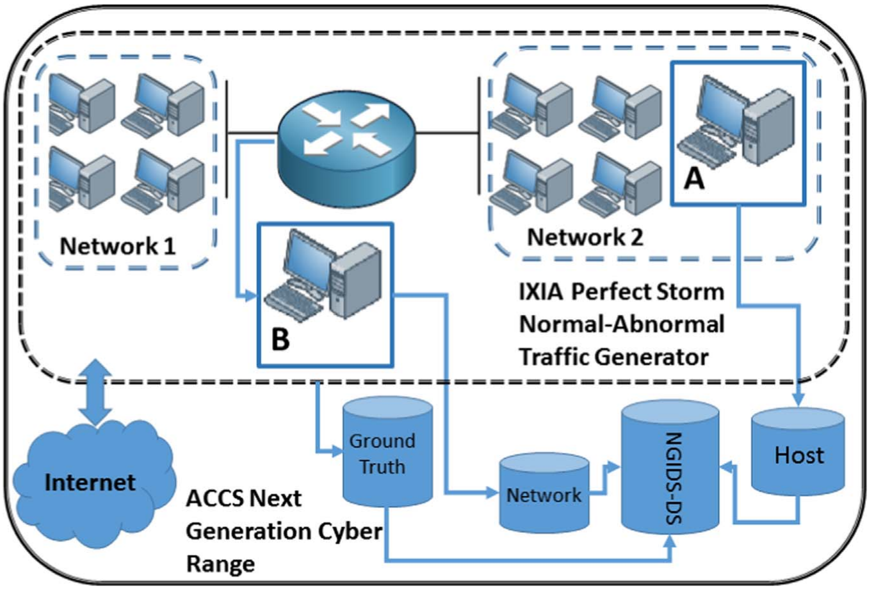

Fig. 1. Testbed architecture.

following required factors pairs: (i) x_2, y_1;(ii) x_3, y_2; (iii) x_4, y_2 and (iv) x_6, y_1. Further, UNSW-NB15 specificity to network logs collection is the only reason of lacking some other required factors pairs. However, it is comparatively better and updated IDS dataset for NIDS research.

It can be observed that the above IDS datasets (excluding UNSW-NB15) lack the required number of factor pairs listed in Tables 1, 2 due to the following four problems: (i) their lack of commercial-level penetration testing resources; (ii) their inability to access real enterprise networks for penetration testing and data collection; (iii) the reality of the dynamics of internet traffic over time; and (iv) their undermining of a careful design process, for instance, ignoring the collection of host and network logs during the simulation.

Unfortunately, UNSW-NB15 also suffers with issue (iv). Further, these issues can affect the design of a general and reliable IDS in terms of accuracy. Hence, it is necessary to generate a realistic IDS dataset with the maximum number of possible factors pairs.

In related work, it is also observed that, while the documentation for most existing IDS datasets includes an explanation of the dataset's generation process, format and quantity, it does not present a metric for evaluating its quality. The authors in Shiravi et al. (2012) suggested incorporating some useful design features similar to those of our crisp input set X elements during the generation of an IDS dataset, on the basis of which they presented a YES/NO-based metric for evaluating its quality. However, it is observed that there are two major problems associated with that metric: (i) It is impossible to assess the actual realistic quality of the implementation of the suggested features (i.e., factors pairs from set X and Y). For instance, the feature of realistic traffic can be achieved in two different ways (i.e., realistically and synthetically) which reflect different qualities: (a) One collects traffic from a real production network and implements the feature of realistic traffic with a maximum quality of realism. (b) One collects traffic from a synthetic network resulting in a low quality of realism when implementing the feature of realistic traffic. (ii) Quantifying and determining the approximate numerical quality of realism of any IDS dataset is not possible using simple YES/NO arguments. Therefore, there is a need for a metric that can handle these problems.

4. NGIDS-DS generation process

To generate the NGIDS-DS with the maximum possible quality of realism, the next-generation cyber range of infrastructure at the Australian Centre for Cyber Security (ACCS), Australian Defence Force Academy (ADFA), Canberra, which was designed according to the guidelines provided in Davis and Magrath (2013), is used. Its key advantage is the availability of the IXIA *Perfect Storm* hardware. The combination of a network traffic-generation appliance and virtual cyber range provides both legitimate traffic and host-based connectivity. The IXIA Perfect Storm tool provides four major capabilities. Firstly, it can produce a mixture of modern normal and abnormal cyber traffic. Secondly, it can generate the maximum number of attacks with different dynamic behaviors based on packs that exploit known common vulnerability exposures (CVE) (Dictionary of attacks, 2016). Thirdly, it can establish profiles of the cyber traffic of multiple enterprises. Fourthly, it can automatically generate the ground truth.

The design of the NGIDS-DS generation process defines four criteria to which the simulation must conform based on related efforts: (i) synthetic realizations of the maximum number of current real-world network attacks, their execution behaviors and normal traffic dynamics; (ii) an automated reporting capacity (i.e., ground truth) that expresses in a relational data format what occurred at what time in terms of identifying the user activity, resource utilization and category of activity (i.e., normal or abnormal); (iii) the capability to capture all the effects of the simulation, i.e., collecting network packets and the operating system logs of hosts while maintaining the performance of the cyber infrastructure; and (iv) the inclusion of attribute x_4 from set X. The purpose of these criteria is to ensure maximum effective rules for achieving the maximum quality of realism during the simulation.

The NGIDS-DS generation environment is presented in Fig. 1. In the pre-simulation phase, inside IXIA *Perfect Storm*, Network 1 is designed to generate a mixture of normal and abnormal traffic as well as operational timings and the profiles of five different enterprises (e-commerce, the military, academia, social media and banks), and Network 2 to act as a collective victim network or critical cyber infrastructure of these enterprises. The abnormal traffic includes seven major attack families: *Exploits; DoS; Worms; Generic; Reconnaissance; Shellcode;* and *Backdoors*. The two capturing machines, i.e., A and B, are designed to ensure that the third criterion is met. Machine A, which has Ubuntu 14.04 and an auditing mechanism installed on it (auditd), is designed to act as an enterprise-critical machine running several services (e.g., storage, Web, FTP, email, NAT, SSH and DNS). Its major purpose is to collect the normal and abnormal effects obtained from the simulation at the level of the victim-host operating system. Machine B, which has Ubuntu 14.04 and tcpdump installed on it, collects network packets originating from Network 1 and moving towards Network 2. Interestingly, in the pre-simulation phase, during the capture of all the operating system's logs at machine A, it is noticed that the machine's performance becomes degraded. To avoid this, the collection mechanisms used on machine A are customized to collect only the data required which include the system's date, time and process id, a system call, an event id and the process's execution paths. The simulation process for developing the NGIDS-DS was begun on March 11, 2016 at 2:45:05 AM and completed on March 16, 2016 at approximately 7:45:12 PM. It was conducted before the simulation to provide 27 h of cyber normal and abnormal traffic scenarios for each of the five different enterprises to be executed.

4.1. Quantitative description of NGIDS-DS

The NGIDS-DS consists of the following five different types of files which can be downloaded at (NGIDS-DS download, 2016): (i) ground-truth.csv; (ii) 99 csv files of host logs; (iii) NGIDS.pcap of the network packets; (iv) feature-descr.csv; and (v) readme.txt.

It can be seen in Table 4 and Fig. 2 that the number of captured network packets is less than that of the host log records, as expected because, for one network event, there would be 82 corresponding host ones (Thompson et al., 1997; Xie et al., 2012). Furthermore, in the host logs, there are fewer attack than normal records, with a ratio of 1:90 which demonstrates the stealthy behaviors of current attack activities (Dictionary of attacks, 2016) in the NGIDS-DS. Moreover, most enterprises now try to invest in protecting their critical cyber infrastructures (Bessani et al., 2008; Mahmood et al., 2010). Therefore, in our testbed, Network 1 replicates the fact that, in a real-world scenario, the threat vectors required to destroy an enterprise-critical network are complex and attacks may come from different entities. Network 2 is the collective synthetic representation of this enterprise-critical network which can be attacked. Therefore, we capture the relevant network packets from the simulation to observe the effects of real threats at the right locations while emphasizing quality over quantity.

4.2. Preliminary analysis of simulation traffic

A preliminary analysis is performed using the traffic features and ground-truth information for the entire simulation period of approximately 4,95,000 s, with the purpose of showing what happens in OSI layers 1, 2, 3, 4 and 7 during the simulation and why. It can be seen in Fig. 3 (top) that the approximate average transmit and receive rates of the ethernet frames are 572 and 455 respectively. This average loss of 117 frames reflects the actual phenomenon of ethernet frames being lost in real-world networks through network noise (Thompson et al., 1997). In Fig. 3 (bottom), in the microseconds unit, it can be seen that the average minimum and frame latencies are 1.2045×10^8 and 1.205×10^8 respectively, with the range of the maximum frame latency from 1.2055×10^8 to 1.20575×10^8. Furthermore, the average instantaneous frame latency is a measure of the sudden jump in the frame latency from the average latency offset which is seen to range from 1.205×10^8 to 1.20599×10^8. While analyzing the ground truth, it is observed that there are many DoS and Distributed DoS (DDoS) attacks in different time ranges, such as 0–48,000, 72,000 to 1,25,000, 1,50,000 to 2,25,000 and up to 4,95,000 s, which orient the pattern of the maximum frame latency. Also, the high number of exploits in the range from 0 to 2,31,000 s constitutes a pattern of average instantaneous frame latency.

Table 4
Summarized quantitative descriptions of NGIDS-DS.

File type	Feature	Quantity
ground-truth csv	Total Records Attributes	313,926 7
99 csv files of host logs	Total Records Attack Records Normal Records Attributes	90,054,160 1,262,426 88,791,734 9
NGIDS.pcap	Total Capture Packets Unique IPs	1,094,231 18

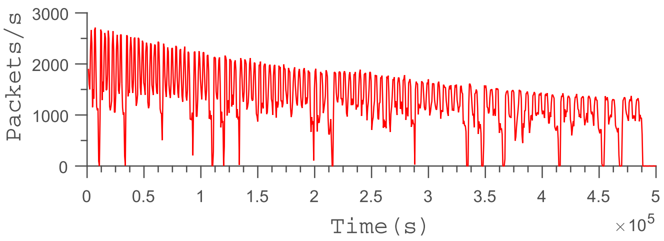

Fig. 2. Quantitative pattern of network packets captured over entire simulation period.

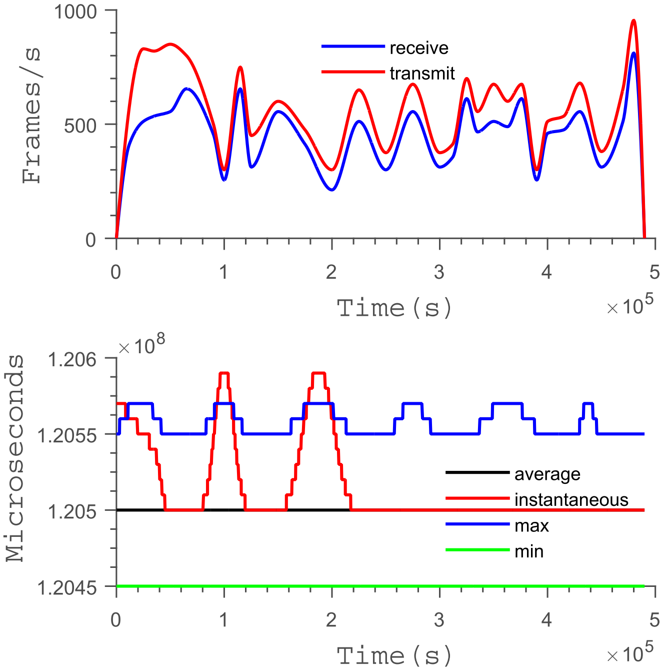

Fig. 3. Transmit/receive rates (top) of ethernet frames and different latencies (bottom) of ethernet frames.

In Fig. 4 (top), in the layer 7 applications, it can be seen that the average response time is 200 ms whereas the effects of the Exploits and densities of the DoS and DDoS types of attacks constitute a pattern of the instantaneous response times of applications, including HTTP, FTP, DNS, SMTP, SSH, IMAPv4, AOLIM, RPCs and Bittorrents, according to the ground truth. Similarly, the average number of the application's attempts and success rates is in the range of approximately 38 to 45 transactions per second without any application failure, as shown in Fig. 4 (bottom). The difference between these rates is obviously due to the time an application requires from an attempt to achieving success.

In the simulation, it is observed that the durations of normal traffic govern the pattern of a client/server establishment rate and those of DDoS attacks that of a client/server closing rate, as shown in Fig. 5 (top). Interestingly, the difference between, and equality of, both rates is a result of the inclusion of a realistic ratio of normal to abnormal traffic as well as the realistic execution of DDoS attacks in the simulation. Moreover, in Fig. 5 (bottom), the concurrent UDP sessions of client/servers are directly proportional to the times of the simulation, as is obvious in the comparison with those of a real network scenario.

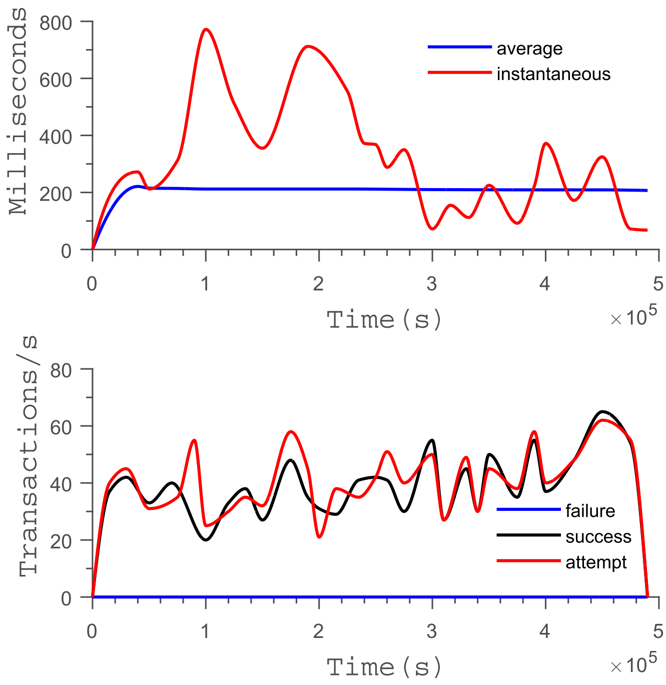

Fig. 4. Different response times (top) and rates (bottom) of layer 7 applications.

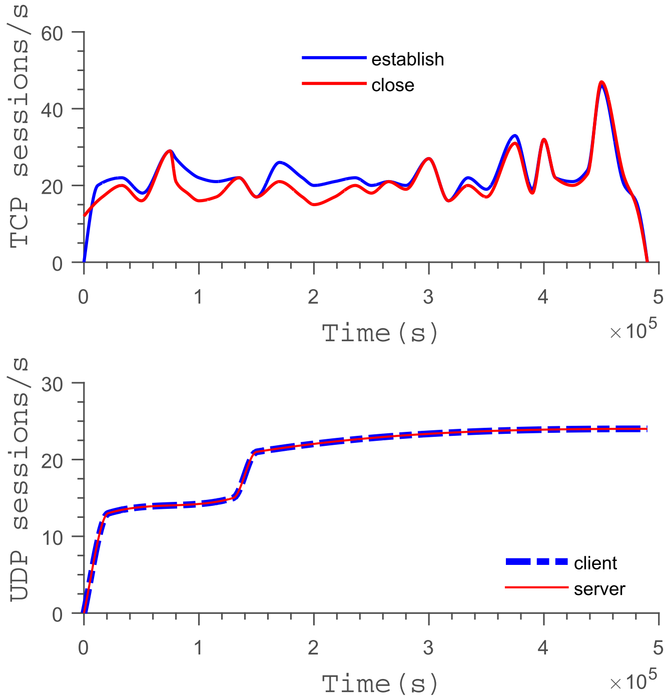

Fig. 5. Different rates (top) and concurrent sessions (bottom) of clients/servers.

5. Evaluating the NGIDS-DS realism comparatively

Due to the design of the NGIDS-DS simulation discussed in Section 3, there are six rules involved (e.g., N=6), (x_1 AND y_1), (x_2 AND y_1), (x_5 AND y_1), (x_6 AND y_1), (x_3 AND y_2) and (x_4 AND y_2). Considering $i = 1 \ldots N$ and the parameters $a = b = c = N$, using Eq. (1), $z_1 = z_2 = z_3 = z_4 = 12.96$ and $z_5 = z_6 = 9.96$ and, similarly, using Eq. (2), $w_1 = w_2 = w_3 = w_4 = 0.16$ and $w_5 = w_6 = 0.08$. Furthermore, the final realism obtained for R is 0.95 and the comparative results shown in Fig. 6 indicate that the quality of realism of the NGIDS-DS is linguistically medium-high and dominant.

There are two main reasons for existing datasets having low and medium qualities of realism: (i) the absence of all the elements of set X in a single IDS dataset; and (ii) the unavailability of a resource-full generation environment to systematically enable the possible quality of realism to be obtained. The answer to the question regarding how the NGIDS-DS achieves a medium-high quality of realism is that, in the generation process, we use a commercial security-tested hardware platform, i.e., IXIA *Perfect Storm*, to realistically represent attribute x_2, and synthetically represent x_3 and x_4 or, in other words, form rules such as (x_2 AND y_1), (x_3 AND y_2) and (x_4 AND y_2). Interestingly, disputes may arise concerning the use of such a commercial tool in the generation process for an IDS dataset and why the NGIDS-DS cannot achieve a high quality of realism. Two justifications are that: (i) it is practically impossible for the low-resource-enabled testbeds used by existing dataset generators to simulate real-world networks, normal traffic dynamics and possible attacks in a short period; and (ii) due to issues regarding integrity, cost and the requirement for uninterrupted business operations, enterprises avoid using their actual production networks for penetrative testing purposes, i.e., obtaining an IDS dataset with high-quality realism is only possible using a real network.

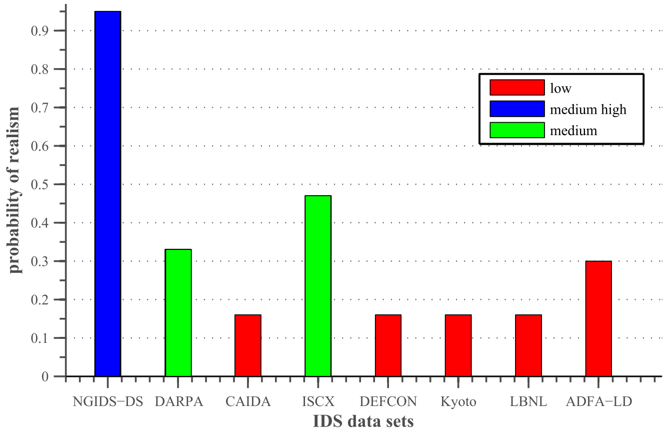

Fig. 6. Comparative analysis of quality of realism of existing datasets.

In order to observe the comparative difference between using the proposed and existing metrics to evaluate the quality of realism of any IDS dataset, consider Tables 3 and 5, and Fig. 6. Table 3 and Fig. 6 show the outcomes of the proposed metric and Table 5 those of the existing one. It is clear that the proposed metric is systematically capable of acquiring the approximate numerical quality of realism of any IDS dataset. On the other hand, as shown in Table 5, it is impossible to assess the actual quantified quality of realism of any IDS dataset due to the lack of the elements of the crisp input set Y, e.g., the ways (synthetic or realistic) of the implementation of the factors of set X. This limitation means that the scores from the actual implementation of the elements of crisp input set X are excluded while quantifying and assessing the quality of realism of an IDS dataset. Moreover, such metric (given in Table 5) does not provide a composite quality indicator for the IDS dataset as a whole.

Table 5
Comparison of existing and proposed datasets using metric format presented in Shiravi et al. (2012) and crisp input set of X elements given in Table 1 (note that NIP means 'no information provided').

Datasets	x_1	x_2	x_3	x_4	x_5	x_6
DARPA (McHugh, 2000)	Yes	No	No	No	NIP	Yes
ISCX (Shiravi et al., 2012)	Yes	No	Yes	No	NIP	Yes
ADFA-LD (Creech and Hu, 2013)	No	No	Yes	No	NIP	Yes
Kyoto (Song et al., 2011)	No	No	No	No	NIP	Yes
DEFCON (Shiravi et al., 2012)	Yes	No	No	No	NIP	No
LBNL, CAIDA (Shiravi et al., 2012)	No	No	No	Yes	NIP	No
NGIDS-DS	Yes	Yes	Yes	Yes	Yes	Yes

6. Conclusion and future work

In this paper, a qualitative modeling approach based on a FLS is used to develop a metric for evaluating the quality of realism of an IDS dataset. It establishes criteria for generating or selecting a realistic dataset for the design of a reliable IDS in future. Also, this metric's capability to influence the generation process of an IDS dataset is demonstrated by the NGIDS-DS which has a medium-high quality of realism and is generated by the intelligent use of the commercial-scale security test hardware platform IXIA *Perfect Storm*. The NGIDS-DS consists of a labeled network and host logs which collectively reflect the current critical cyber infrastructures of different enterprises in both normal and abnormal scenarios. The preliminary analysis of the simulation traffic shows that the realistic attack behaviors and normal traffic dynamics of real-world networks are included in the NGIDS-DS. In future, first more investigations regarding ground-truth information and captured data will be carried out to determine useful features for future IDS designs. Second, the machine learning and data mining models will be explored to assess the complexity of the NGIDS-DS.

Acknowledgements

This project is supported by ARC grants LP110100602, LP100200538, LP100100404 and Y. Xie's work was supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2014A030313130).

References

Ahmed, M., Mahmood, A.N., Hu, J., 2016. A survey of network anomaly detection techniques. J. Netw. Comput. Appl. 60, 19–31.

Bessani, A.N., Sousa, P., Correia, M., Neves, N.F., Verissimo, P., 2008. The crutial way of critical infrastructure protection. IEEE Secur. Priv. 6 (6), 44–51.

Creech, G., Hu, J., 2013. Generation of a new ids test dataset: Time to retire the kdd collection. In: 2013 IEEE Wireless Communications and Networking Conference (WCNC), pp. 4487–4492.

Creech, 2014. Developing a high-accuracy cross platform host-based intrusion detection system capable of reliably detecting zero-day attacks. (Ph.D. thesis). University of New South Wales.

Davis, J., Magrath, S., 2013. A survey of cyber ranges and testbeds. Tech. Rep., DTIC Document.

Dictionary of attacks, 2016. ⟨https://cve.mitre.org/⟩ (accessed 30 July 2016)).

Gogoi, P., Bhuyan, M.H., Bhattacharyya, D., Kalita, J.K., 2012. Packet and flow based network intrusion dataset. In: International Conference on Contemporary Computing. Springer, pp. 322–334.

Haider, W., Hu, J., Yu, X., Xie, Y., 2015. Integer data zero-watermark assisted system calls abstraction and normalization for host based anomaly detection systems. In:Proceedings of the IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), 2015, pp. 349–355.

Haider, W., Hu, J., Xie, M., 2015. Towards reliable data feature retrieval and decision engine in host-based anomaly detection systems. In: Proceedings of the IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), pp. 513–517.

Haider, W., Creech, G., Xie, Y., Hu, J., 2016. Windows based data sets for evaluation of robustness of host based intrusion detection systems (ids) to zero-day and stealth attacks. Future Internet 8 (3), 29.

Hoang, X.D., Hu, J., Bertok, P., 2009. A program-based anomaly intrusion detection scheme using multiple detection engines and fuzzy inference. J. Netw. Comput. Appl. 32 (6), 1219–1228.

Hu, J., Khalil, I., Han, S., Mahmood, A., 2011. Seamless integration of dependability and security concepts in soa: a feedback control system based framework and taxonomy. J. Netw. Comput. Appl. 34 (4), 1150–1159.

Mahmood, A.N., Hu, J., Tari, Z., Leckie, C., 2010. Critical infrastructure protection: resource efficient sampling to improve detection of less frequent patterns in network traffic. J. Netw. Comput. Appl. 33 (4), 491–502.

McHugh, J., 2000. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. Inf. Syst. Secur. (TISSEC) 3 (4), 262–294.

Moustafa, N., Slay, J., 2015. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set) In: Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, pp. 1–6.

Moustafa, N., Slay, J., 2016. 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.: a Glob. Perspect. 25 (1–3), 18–31.

NGIDS-DS download, 2016. ⟨https://research.unsw.edu.au/people/professor-jiankun-hu⟩, (accessed 30 July 2016)).

Penetration Tester, 2016. ⟨https://strikecenter.ixiacom.com/docs/BPS_UserGuide_3.5.pdf⟩ (accessed 30 July 2016).

Sangster, B., O′Connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., Conti, G.J., 2009. Toward instrumenting network warfare competitions to generate labeled datasets. In: CSET.

Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A., 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput. Secur. 31 (3), 357–374.

Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., Nakao, K., 2011. Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation. In: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security. ACM, pp. 29–36.

Sugeno, M., Yasukawa, T., 1993. A fuzzy-logic-based approach to qualitative modeling. IEEE Trans. Fuzzy Syst. 1 (1), 7–31.

Thompson, K., Miller, G.J., Wilder, R., 1997. Wide-area internet traffic patterns and characteristics. IEEE Netw. 11 (6), 10–23.

Vasudevan, A., Harshini, E., Selvakumar, S., 2011. Ssenet-2011: a network intrusion detection system dataset and its comparison with kdd cup 99 dataset. In: 2011 Proceedings of the Second Asian Himalayas International Conference on Internet (AH-ICI). IEEE, pp. 1–5.

Xie, Y., Hu, J., Tang, S., Huang, X., 2012. A structural approach for modelling the hierarchical dynamic process of web workload in a large-scale campus network. J. Netw. Comput. Appl. 35 (6), 2081–2091.

Zuech, R., Khoshgoftaar, T.M., Seliya, N., Najafabadi, M.M., Kemp, C., 2015. A new intrusion detection benchmarking system. In: FLAIRS Conference. pp. 252–256.

Waqas Haider received the BS and MS degrees from COMSATS institute of information of information technology and IQRA university Pakistan respectively. He is currently a Ph.D. student at the School of Engineering and Information Technology in University of New South Wales at Australian Defence force Academy, Canberra. His research interests include network and host based intrusion detection systems.

Jiankun Hu is a full professor of Cyber Security at the School of Engineering and Information Technology, the University of New South Wales at the Australian Defence Force Academy (UNSW@ADFA), Australia. His major research interest is in computer networking and computer security, especially biometric security. He has been awarded seven Australia Research Council Grants. He served as Security Symposium Co-Chair for IEEE GLOBECOM '08 and IEEE ICC '09. He was Program Co-Chair of the 2008 International Symposium on Computer Science and its Applications. He served and is serving as an Associate Editor of the following journals: Journal of Network and Computer Applications, Elsevier; Journal of Security and Communication Networks, Wiley; and Journal of Wireless Communication and Mobile Computing, Wiley. He is the leading Guest Editor of a 2009 special issue on biometric security for mobile computing, Journal of Security and Communication Networks, Wiley.

He received a Bachelor's degree in industrial automation in 1983 from Hunan University, PR China, a Ph.D. degree in engineering in 1993 from the Harbin Institute of Technology, PR China, and a Master's degree for research in computer science and software engineering from Monash University, Australia, in 2000.

In 1995 he completed his postdoctoral fellow work in the Department of Electrical and Electronic Engineering, Harbin Shipbuilding College, PR China. He was a research fellow of the Alexander von Humboldt Foundation in the Department of Electrical and Electronic Engineering, Ruhr University, Germany, during 1995–1997. He worked as a research fellow in the Department of Electrical and Electronic Engineering, Delft University of Technology, the Netherlands, in 1997. Before he moved to RMIT University Australia, he was a research fellow in the Department of Electrical and Electronic Engineering, University of Melbourne, Australia.

Jill Slay is Professor of Cyber Security and Director of the Australian Centre for Cyber Security at UNSW Canberra @ ADFA. This centre has developed critical mass in cross-disciplinary research and teaching in Cyber Security to serve the Australian Government and Defence Force and help strengthen the Digital Economy.

Dr Benjamin Turnbull is a Senior Lecturer with the Australian Centre for Cyber Security, at the University of New South Wales. His research interests include cyber-resilience, cyber-kinetic impact analysis and novel methods for network analysis. He previously worked for the Australian Defence Force.

Yi Xie received the B.S. degree, M.S. degree and Ph.D. degree from Sun Yat-Sen University, Guangzhou, China. He was a Visiting Scholar with George Mason University and Deakin University during 2007–2008, and during 2014–2015, respectively. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-Sen University. His research interests focus on network security and user behaviour.