Towards generic quality assessment of synthetic traffic for evaluating intrusion detection systems

Adrien Schoen, Gregory Blanc, Pierre-François Gimenez, Yufei Han, Frédéric Majorczyk, Ludovic Mé

Towards generic quality assessment of synthetic traffic for evaluating intrusion detection systems

Adrien Schoen*, Gregory Blanc§, Pierre-François Gimenez†, Yufei Han*, Frédéric Majorczyk‡, Ludovic Mé*
*Inria, Univ. Rennes, IRISA, {firstname.lastname}@inria.com
†CentraleSupélec, Univ. Rennes, IRISA, pierre-francois.gimenez@centralesupelec.fr
‡DGA-MI, Univ. Rennes, IRISA, frederic.majorczyk@intradef.gouv.fr
§Samovar, Télécom SudParis, Institut Polytechnique de Paris, gregory.blanc@telecom-sudparis.eu

***Abstract*—Network Intrusion Detection Systems (NIDSes) evaluation requires background traffic. However, real background traffic is hard to collect. We hence rely on synthetic traffic generated especially for this task. The quality of the generated traffic has to be evaluated according to some clearly defined criteria. In this paper, we show how to adapt the quality assessment solutions proposed for different fields of data generation such as image or text generation to network traffic. We summarize our study by discussing the criteria that evaluate the quality of a generated network traffic and by proposing functions to evaluate these criteria. This is the first contribution in the context of the Ph.D. thesis of Adrien Schoen.**

***Index Terms*—Intrusion Detection, Synthetic traffic, Evaluation, Data generation**

I. Introduction

Damages caused by sophisticated cyber-attacks have increased significantly during recent years. It is necessary to provide accurate and fast detection of cyber-attacks against key IT/OT infrastructures. Network Intrusion Detection Systems (NIDS) monitors network traffic to identify malicious activities. Nowadays, these tools have become crucial for the security of key infrastructures. Therefore, the rigorous and complete evaluation of their capabilities is a key issue. The essential capabilities of these tools are 1) to produce an alert for any attack (no false negative) and 2) to produce an alert only in case of an attack (no false positive). Thus, evaluating NIDS consists of providing them with network traffic to verify whether these two capabilities are met.

To this end, it must be provided with legitimate traffic (to verify that it does not generate false alerts) and intrusive traffic (to verify that it detects all attacks in this traffic). In this paper, we focus on legitimate traffic. There are two ways to collect benign network traffic [8]: public network traffic datasets (often simulated) and records of network traffic in private network infrastructures. The former solution suffers from becoming obsolete quickly, as the network traffic to analyze evolves as quickly as usages, technologies, and network protocols. The latter solution requires time-consuming data labeling process. Moreover, some of the recorded data might be sensitive.

For this reason, we focus on an alternative data source that took advantage of the generative deep learning techniques developed in recent years: synthetic Network Traffic Generation [2,3,8]. This approach allows generating a large quantity of legitimate and clean data. However, as pointed out by several pieces of work [2,6], the main drawback of network traffic generation is that no generally applicable evaluation method is available to measure the quality of generated network traffic.

In Section II, we present a brief overview of the criteria defined for several other domains (Subsection II-A) and explain how scoring functions (Subsection II-B) can evaluate these criteria. Finally, we study how to adapt these scoring functions to the case of network traffic generation (Section III). Section IV concludes the article.

This work has been partly realized thanks to a doctoral grant from the cyber excellence pole (PEC: DGA, Brittany Region).

II. Quality evaluation of generated data

A. Criteria for quality evaluation

The quality evaluation criteria correspond to the very general properties of the generated data. Most of the published evaluation criteria [1,4,9] are specific to some application fields. Besides, [6] states that generated traffic has to be evaluated in regards to its final usage. Consequently, there is no generally applicable criterion proposed for this domain. Nevertheless, three of them appear to be reusable in principle to various types of data:

- **Realism** (also called fidelity): a synthetic sample should be sampled from the same distribution as the real data.
- **Diversity** (also called fairness): the distribution of the generated samples should have the same variability as the real data.
- **Originality** (also called authenticity): a generated sample should be sufficiently different from the samples of the real distribution.

Note that originality and realism can be mutually exclusive. For example, if a generated sample is just a copy of a real sample, it will be realist by construction but not original. Therefore a trade-off is necessary.

B. Scoring functions for quality evaluation

The criteria defined in the previous subsection are assessed through scoring functions that output scores. A score can be computed for a sample or an entire distribution. A sample-level scoring function produces a score for each sample.

Various scoring functions have been proposed to evaluate one or more of the above criteria in specific fields, i.e.,

tabular data generation, image generation, and text generation. For the sake of brevity, we only present the most diverse and representative scoring functions. We also present scoring functions specifically created for assessing the quality of a generated network traffic. The characteristics of these scoring functions are presented in Table I.

1) Tabular data generation: This research domain deals with the generation of numerical data in tabular series, such as ones used for many Big Data applications [1]. The quality of generated tabular data is often evaluated through numerical metrics. For example, "recall" evaluates the data diversity and "precision" its realism. Other scoring functions have been proposed, such as Coverage, Density, and Authenticity [1].

2) Image generation: To evaluate the quality of generated images, various scoring functions have been proposed, but the most popular are *Inception Score* and *Frechet Inception Distance* (FID), with their multiple adaptations, such as *Memorization Informed FID* (MiFID) [1].

3) Text generation: The quality of generated texts can be evaluated considering sequences of n words tuples called n-grams [9]. For example, BLEU [7] measures the proportion of the n-grams of a generated text that exist in the real-world reference documents. It evaluates the realism of the generated text. Another scoring function *WMD* (Word Mover's Distance) [5] is based on the distance between the n-gram frequencies of the real and generated texts to evaluate realism. It can also be used to assess diversity by computing the distance between two generated texts. Other scoring functions include ROUGE, SelfBLEU or, more recently, machine-learning-based scoring functions such as BERTscore [10].

4) Network traffic generation: Various scoring functions (or actually scoring "processes") have been proposed [2,3,6,11] but none has become well established among the community. The evaluation of the quality of a generated network traffic implies, on the one hand, to evaluate the quality of each of the generated packets, and on the other hand, to evaluate the quality of the sequence of these packets. These two aspects are indeed defined by network protocols (respectively by frame format and protocol state machine). In addition, in some cases, the generated data takes the form of a description of network flows (number of packets, duration, inter-packet arrival time, etc.) such as the ones produced by NetFlow or CICFlowMeter.

The literature proposes scoring functions tailored to evaluate the quality of individual packets or traffic flow descriptions. However, to our best knowledge, no scoring function has been designed for the quality evaluation of a sequence of packets.

Two scoring functions have been proposed to assess the quality of an individual packet. PcapGAN visualization test [3] consists in verifying whether a packet analyzer tool (e.g., Wireshark) can parse the generated traffic without error. In the PAC-GAN quality test [2], the generated packets are sent over the network and the answer (if any) is analyzed.

Several scoring functions assess the quality of flow descriptions. The GANvsReal score (GvR score) [11] evaluates the realism and diversity of the generated traffic. It is computed from the difference of accuracy between a classifier trained on the real data and a classifier trained on the synthetic data: if the generated traffic is realistic and diverse, the difference of accuracy is low. However, this scoring function cannot be directly applied in our case, since its classification task needs benign and malicious network traffic. Domain Knowledge Check (DKC) [8] verifies seven rules, such as verifying that the destination port (e.g., 80 or 443) and the transport protocol (e.g., TCP) are consistent.

TABLE I
SCORING FUNCTIONS SUMMARY. "R", "D", "O" AND "C" STAND FOR REALISM, DIVERSITY, ORIGINALITY AND COMPLIANCE RESPECTIVELY.

Data type	Scoring functions	Input	R	D	O	C
Tabular	Recall, Density	Distribution		✓		
	Precision, Coverage	Distribution	✓			
	Authenticity	Sample			✓	
Image	Inception Score, FID	Distribution	✓	✓		
	MiFID	Distribution	✓	✓	✓	
Textual	BLEU, ROUGE	Sample	✓			
	WMD	Sample	✓	✓		
	BERTscore	Distribution	✓	✓		
	SelfBLEU	Distribution		✓		
Network traffic	DKC, PcapGAN test	Sample				✓
	PAC-GAN test	Sample				✓
	GvR	Distribution	✓	✓		

This last scoring function does not directly evaluate any of the three criteria introduced earlier. It shows that the evaluation of the quality of a generated network traffic cannot be limited to the assessment of these three criteria. The references [2,8] notice that even a very small difference between a generated sample and a real one can lead to unusable traffic. For this reason, we introduce a new criterion in the next Section: the compliance to network specifications.

III. CONTRIBUTIONS ON GENERATED NETWORK TRAFFIC EVALUATION

We need to propose an evaluation method that takes into consideration the three criteria presented in subsection II-A (realism, diversity, originality) and an additional criterion (compliance) we introduce in this section. In the following, we study whether the various scoring functions presented above can be adapted to network traffic generation.

A. Compliance of generated traffic

Beyond realism, diversity, and originality, a generated network traffic must also conform to specifications: the order of the fields, the acceptable values in these fields, and the possible sequences of packets are rigorously specified. Although the idea that generated traffic must conform to the standards is of course present in the literature, no specific criterion is defined to convey this idea. To some extent, compliance and realism are related criteria. However, they are distinct: realism assesses how a synthetic sample fits in the reference data distribution while compliance assesses the validity of a sample with respect to the specifications of network protocols.

B. Scoring functions for network traffic generation

Compliance can be easily evaluated through the scoring functions presented in Subsection II-B4. What function to

use depends on the type of network data to be evaluated. For a flow description, Domain Knowledge Check is the only scoring function that can be used. If the generation outputs network packets directly, then the PcapGAN visualization is preferable to the PAC-GAN quality test because it does not involve sending the packet out. No network-specific scoring function is tailored to packets sequence evaluation. However, such compliance is implicitly verified by scoring functions that work on traffic flow descriptions: indeed, if the packet sequence is not compliant, no flow description can be produced.

1) Relevance of tabular data generation scoring functions: Packets headers (whose size is constant for a fixed protocol) or characteristics of network flows (number of packets, duration of the flow, transport protocol, etc.) can be easily expressed as tabular data. Thus, the scoring functions of Subsection II-B1 can be directly used.

2) Relevance of image generation scoring functions: The scoring functions for image generation presented in Section II-B requires the use of an existing classifier, typically Inceptionv3, for a classification task. However, we do not expect such scoring functions to be easily adaptable to the context of network traffic generation. Indeed, while Inceptionv3 has to learn high-level concepts to classify images, a network classifier could trivially verify certain bytes to predict the network protocol of a packet, for example. This would make the Inception-Score-based scoring functions trivial: to be considered realist, a packet would only need to construct a few bytes correctly. For this reason, the real challenge of adapting these scoring functions is not building the classifier itself, but choosing the right, non-trivial classification task.

3) Relevance of text generation scoring functions: Text generation has some similarities with network traffic generation: both produce sequential data. While text generation scoring functions are based on n-grams of tokens, they can be easily adapted to work on n-grams of bytes, e.g., for evaluating generated packet headers. They might also be adapted to n-grams of packets, although it would not be straightforward: the difficulty is to transform the packets into relevant tokens. For example, one could choose to merge any ICMP packet into the same token, or differentiate tokens depending on the values of the relevant fields. Other scoring functions of this domain rely on machine learning methods (e.g., BERTscore). They will be difficult to adapt because it depends on a pre-trained word embedding model. To the best of our knowledge, there is not such model for network traffic yet. Training one can require a large amount of data and considerable efforts. Therefore, those scoring functions should be avoided.

As a conclusion, this section shows that 1) network-specific scoring functions can be used to evaluate the compliance of packets, packet sequences and network flow descriptions, 2) tabular scoring functions can be directly applied to the packet headers and network flow descriptions evaluation to evaluate their realism, originality and diversity, 3) text scoring functions could be adapted to sequences of packets to evaluate their realism, originality and diversity, although not in a straightforward manner, and 4) image scoring functions would probably be very challenging to adapt correctly.

IV. Conclusion

In this paper, we reviewed the state of the art in data generation and highlighted some evaluation criteria in non-network related domains. We proposed the compliance to network protocols as a novel evaluation criterion for synthesized network traffic. Indeed, the space of valid network packets is far more constrained than the ones in the domain of images or texts: even one bit flip can make a packet non-compliant.

Currently, network-specific scoring functions do not take into account all the aspects of network data. In particular, they do not take into account the originality criteria. In addition, they generally work either at the packet level or at the flow description level. However, they cannot handle useful properties relative to a *sequence* of packets.

To this end, we verified whether scoring functions from other domains could be applied or adapted to network traffic generation. We conclude that even though network-specific scoring functions should be picked in priority when possible, none at the moment can be used to evaluate originality or diversity. We recommend to adapt scoring functions defined for tabular or text generation to evaluate these criteria. Moreover, the existing scoring functions do not apply to the packets sequence level. This opens the research avenues that we have discussed.

References

[1] Borji, A. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding 215* (2022), 103329.

[2] Cheng, A. Pac-gan: Packet generation of network traffic using generative adversarial networks. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (2019), pp. 0728–0734.

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

[4] Gulrajani, I., Raffel, C., and Metz, L. Towards GAN benchmarks which require generalization. In *International Conference on Learning Representations* (2019).

[5] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (2015), ICML'15, JMLR.org, p. 957–966.

[6] Molnár, S., Megyesi, P., and Szabó, G. How to validate traffic generators? In *2013 IEEE International Conference on Communications Workshops (ICC)* (2013), pp. 1340–1344.

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, USA, July 2002), Association for Computational Linguistics, pp. 311–318.

[8] Ring, M., Schlör, D., Landes, D., and Hotho, A. Flow-based network traffic generation using generative adversarial networks. *Computers & Security 82* (2019), 156–172.

[9] Sai, A. B., Mohankumar, A. K., and Khapra, M. M. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv. 55*, 2 (jan 2022).

[10] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations* (2020).

[11] Zingo, P., and Novocin, A. Can gan-generated network traffic be used to train traffic anomaly classifiers? In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (2020), pp. 0540–0545.