Research on a Generative Adversarial Network Based Framework for Cyber Training Network Generation

Dong-Wook KiM[1], Gun-Yoon-Shin[1], Younghoan Jang[1], Seungjae Cho[2], Kwangsoo Kim[2], Jaesik Kang[2] and Myung-Mook Han[1*]

[1] Department of AI Software, Gachon University, Seongnam-si, Republic of Korea.
{kog7306, jang0h}@naver.com, tlsrjsdbs@gmail.com,
mmhan@gachon.ac.kr
[2] Cyber Electronic Warfare R&D, LIG Nex1, Seongnam-si, Republic of Korea.
{Seungjae.cho, jaesik.kang, kwangsoo.kim}@lignex1.com

Abstract

Cybersecurity is a crucial discipline that centers around safeguarding computer and network systems against cyber threats. It is imperative to provide education and training to tackle real-world cyber-attacks. Cyber training environments cater to this requisite by simulating diverse cyber-attack and defense scenarios. However, creating and executing tests, experiments, and training sessions for this objective demands immense resources and effort. This study proposes a new approach to boost the effectiveness of cybersecurity training, with a specific emphasis on applying GAN models to optimize network topology. Our objective is to enhance the diversity and realism of training.

Keywords: Cyber training, Network simulation, Network topology, GAN

1 Introduction

Cybersecurity is a crucial discipline that seeks to protect computer and network systems from online threats. Institutions worldwide recognize the importance of cybersecurity education and training to counter these risks. However, theoretical knowledge alone cannot fully secure systems in real-world settings. Enhancing response capabilities during actual scenarios is urgently needed.

The 7th International Conference on Mobile Internet Security (MobiSec'23), Dec. 19-21, 2023, Okinawa, Japan, Article No. 65

*Corresponding author: Department of AI Software, Gachon University, Seongnam-si, 13120, Republic of Korea, mmhan@gachon.ac.kr

Cybersecurity training environments offer a platform for simulating diverse cyber-attack and defense scenarios. Nevertheless, the development and implementation of tests, experiments, and training sessions in such environments require considerable resources and effort, mainly due to the need for highly skilled professionals to ensure the training's effectiveness and precision.

The objective of Cyber Training is to cultivate skilled professionals and enhance their capacity to address diverse cyber threats promptly and competently. Such proficiency empowers institutions and organizations to proficiently mitigate the constant and advancing jeopardy of cybercriminals. The simulation of realistic attack scenarios and diverse situations enables personnel to enhance their technical skills and continually improve their abilities to solve problems, work together, and make decisions. This training plays a vital role in boosting their capacity to respond effectively to urgent situations in the field. Simultaneously, it helps in creating a more robust security framework for professionals in the cybersecurity field.

The aim of this research is to propose a new approach to improve the efficiency and scope of cybersecurity training. To address the constraints of repetitive training scenarios and to ensure trainees are prepared to respond to diverse and realistic cyber threats, we present a methodology that generates arbitrary network topologies and utilizes the GAN(Generative Adversarial Networks) model for optimization. Our methodology aims to provide trainees with diverse experiences, preparing them for a wide range of cyber threat situations. To address the constraints of repetitive training scenarios and to ensure trainees are prepared to respond to diverse and realistic cyber threats, we present a methodology that generates arbitrary network topologies and utilizes the GAN model for optimization.

Our methodology covers a range of processes spanning from creating network topologies to learning the GAN model, assessing traffic distribution performance, and refining the model and algorithm. The proposed methodology is set to enhance diversity and realism in cyber security training. The proposed approach is expected to significantly improve trainees' response capabilities and problem-solving skills, thereby enhancing the efficacy and efficiency of cyber security training.

2 Related Work

Constructing a cyber security training environment is a complex process where mistakes can occur due to the prevalence of manual tasks. Therefore, numerous researchers devote their attention to designing and building an infrastructure that can lead to more efficient cyber security training. To meet the demands of the intricate nature of cyber security, which requires extensive work in varied scenario modeling and network traffic creation, automation methodologies are currently being researched. Studies are currently being conducted utilizing Software Defined Networking (SDN), a programmable network technique, for network simulation (Keshari, Kansal, & Kumar, 2021).

In programmable network operation environments, efforts to automate the generation of realistic data using semi-supervised methods of Generative Adversarial Deep Neural Networks are being made to address security vulnerabilities of SDN (Ahmed & George, 2020). Additionally, there is an increasing interest in using deep learning and generative models to produce realistic synthetic Wide Area Networks (WAN) (Dietz, Seufert, & Hoßfeld, 2022). Artificial intelligence models can enhance technology to automate cyber training. An automated cyber security training environment guarantees consistent training, reduces human intervention during the training period, and simultaneously lowers costs while maximizing efficiency.

Therefore, researchers are exploring software-defined networking and artificial intelligence technologies to enhance the efficiency of cybersecurity education by automating cybersecurity training environments. This approach leads to a significant reduction of costs and increase in effectiveness. Therefore, researchers are exploring software-defined networking and artificial intelligence technologies to enhance the efficiency of cybersecurity education by automating cybersecurity training environments.

3 Network Topology Framework Based on Generative Adversarial Neural Networks

In this section, we introduce the step-by-step methodology for automatically generating network topology based on generative adversarial neural networks to provide a variety of environments by creating arbitrary network topologies, aiming to minimize repetitive cyber training networks.

3.1 Network topology and GAN

Based on an arbitrary network topology, during the distribution of actual traffic for generating similar topologies, the structure of Generative Adversarial Networks (GAN) must satisfy the key criteria between the generator $G(z)$ and the discriminator $D(x)$ according to the min-max game. The generator utilizes topology information input based on an arbitrary noise distribution $P_x(z)$ to create a new network topology. In contrast, the discriminator evaluates the traffic distribution efficiency of this created network topology and compares it with the traffic distribution of the actual topology. During the learning process, the discriminator D aims to correctly identify the actual data by maximizing $logD(x)$ and to correctly reject the generated data by maximizing $log(1\text{-}D(G(z)))$. The generator G, to deceive the discriminator, maximizes $D(G(z))$, which is equivalent to minimizing $log(1\text{-}D(G(z)))$.

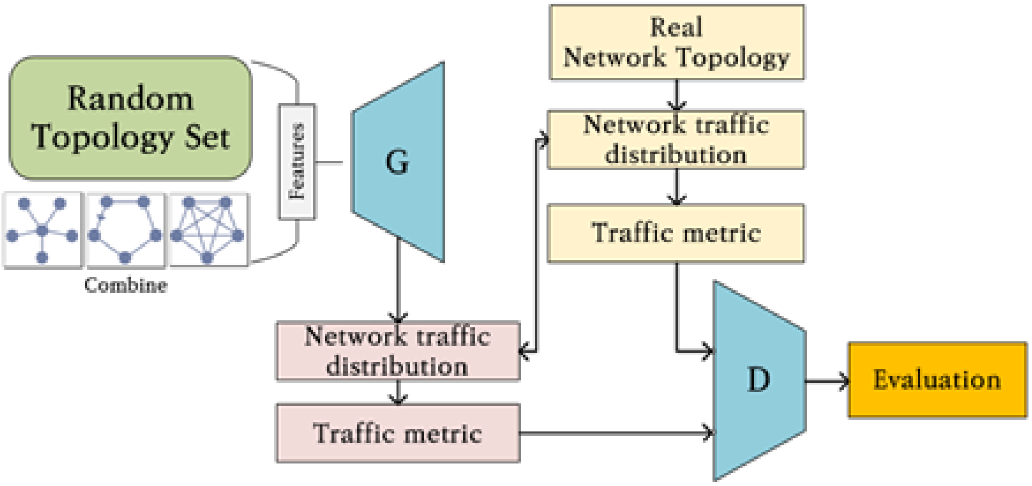

Figure 1 : Proposed Framework

According to this functionality, Figure 1, illustrates that initially, the generator provides input values of the characteristics of topologies of various sizes connected through random graph models, such as bus, star, ring, mesh, and tree topologies, for constructing an arbitrary network topology. The random graph model can utilize the Erdős-Rényi (ER) model (Lima, Sousa, & Sumuor, 2008) and the Barabási-Albert (BA) model (Su, et al., 2014), and for regular topologies (Xia, Fan, & Hill, 2010) can be applied. In this regard, the discriminator needs to be provided input values of the same size (nodes, edges) as the actual network topology to differentiate. The generative model can learn the data distribution according to the characteristics of the input topology and provide approximate data. Since it's difficult to deceive the discriminator with a topology generated with approximate characteristics, the generator and discriminator are designed to balance according to the characteristics of the distributed traffic.

3.2 Network Performance Metrics

When a similar network topology graph is generated in the creation model, traffic is distributed across the topology to measure the amount of traffic occurring throughout the network. During this process, the metrics measured include Latency (delay time) to gauge the quality of network communication, delay time between packet transmissions, packet loss, and error rate to evaluate the network's performance and reliability. The equations for each evaluation metric are as follows.

- $Latency = t_{destination} - t_{source}$
- $Jitter = PacketDelayVariation(PDV)$
- $Packet\ Loss\ Rate = \frac{Number\ of\ Lost\ Packet}{Total\ Sent\ packet} \times 100\%$
- $Error\ Rate = \frac{Number\ of\ Error\ Bits}{Total\ Sent\ Bits} \times 100\%$

In network communication, it is crucial to comprehend and measure performance objectively (Vasilev, Leguay, Paris, Maggi, & Debbah, 2018). An indispensable measure of such performance is Latency, which quantifies the time duration for a packet to journey from source to destination. Essentially, lower latency denotes faster data transit, which is of critical importance in real-time applications, where even slight delays can have deleterious effects. The complementary concept of Jitter is also noteworthy. Jitter measures the inconsistency of packet delay when received at the destination. It results from variance in time within packets arriving, caused by network congestion, timing drift, or other network anomalies. The Packet Loss Rate is another essential metric that indicates the network's reliability. This metric demonstrates the percentage of packets that fail to reach their intended destination. Elevated packet loss rates may indicate network congestion, signal degradation, or other anomalies that could negatively affect the user experience. Finally, the Error Rate provides information on the data's integrity during transmission. This metric measures the portion of transmitted bits that were received with errors. Increased error rates could indicate degraded link quality, which often leads to retransmissions and reduced network efficiency.

These metrics are essential in assessing network performance and reliability. Depending on the network environment and its needs, various Quality of Service (QoS) indicators may be critical. The discriminator utilizes these metrics' attributes to compare and classify bandwidth on a real-life network topology, in line with the traffic flowing through the available network topology. Through this categorization, we implement structures that are appropriate for our instructional network situations.

3.3 Network topology Embedding

Network topology embedding aims to represent graph representations as low-dimensional vectors. This enables prediction tasks, node classification, knowledge graph representation, visualization, clustering, etc. through network analysis. For network topology embedding, there are node embedding and edge embedding methods based on nodes and edges for graph features (Wang, et al., 2022). In this section, we present only node embeddings for network topology embeddings. Node embedding is a process in which each node of a given graph is passed through an arbitrary encoder and transformed into a vector located in the embedding space. The goal is to encode nodes so that their similarity in the embedding space is like their similarity in the original graph network. There are approaches based on adjacency, distance, path, and random walk, and the random walk of node2vec (Grover & Leskovec, 2016) is well known (Zhou, Liu, Liu, Liu, & Gao, 2017).

To provide input to the generation model of a GAN via graph embedding, we need to provide a realistic distribution of data based on the distribution of connections between nodes in the graph. To do this, we provide a sample of similar nodes by generating all node pairs that are within a given window size (the range of neighboring nodes to consider when generating node pairs) on a given path. We define the node pair generation process as follows.

For a given path $P= \{p_1,p_2,\ldots,p_n\}$, pairs of nodes within a window size w are generated. Specifically, for each node p_i, all pairs of p_i and p_j are generated, where j ranges from $max(i\text{-}w,1)$ to $min(i+w, n)$. This can be expressed as:

$$N(P,w)=\bigcup_{i=1}^{n}\{(p_i,p_j)|j\in[\max(i\text{-}w,1),\min(i+w,n)]\wedge i\neq j\}$$

represents the set of all node pairs generated for a given path P and window size w, creating pairs of each node p_i with other nodes p_j within the window range. These generated pairs are used as the input distribution for the generation model, where the generator loss measures how similar the generated data is to the actual data, and the discriminator evaluates its performance in link prediction by distinguishing how well each node pair is a real or generated connection.

4 Conclusion

This study aims to reduce the resources required for repetitive cyber training systems and establish an effective cybersecurity framework. Creating varied network environments for cyber training and conducting simulations by circulating traffic can be expensive and time-consuming. Nonetheless, we seek to present a unique network training scenario using the GAN model. The GAN model can approximate current network topologies, providing significant help in constructing simulation environments without overtaxing resources for actual network settings. Moreover, the method stated in this research incorporates the diverse features and traffic patterns of tangible networks, thus enhancing the efficiency and realism of cybersecurity training. Future research will utilize GAN-based techniques for generating network topologies in diverse environments and scenarios, with the aim of further investigating their effectiveness and efficiency. The expected results of this study will serve as a critical basis for improving the quality of cyber security training.

Acknowledgement

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) grant funded by the Korea government's Defense Acquisition Program Administration (DAPA) (No. KRIT-CT-21-037), and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (RS-2023-00211871).

References

Ahmed, A., & George, K. (2020). SDN-GAN: Generative Adversarial Deep NNs for Synthesizing Cyber Attacks on Software Defined Networks. *On the Move to Meaningful Internet Systems: OTM 2019 Workshops*, 211-220. doi:10.1007/978-3-030-40907-4_23

Dietz, K., Seufert, M., & Hoßfeld, T. (2022). Comparing Traditional and GAN-based Approaches for the Synthesis of Wide Area Network Topologies. *2022 18th International Conference on Network and Service Management (CNSM)*, 64-72. doi:10.23919/CNSM55787.2022.9964866

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855-864.

Keshari, S. K., Kansal, V., & Kumar, S. (2021). A Systematic Review of Quality of Services (QoS) in Software Defined Networking (SDN). *Wireless Pers Commun 116*, 2593-2614. doi:10.1007/s11277-020-07812-2

Lima, F., Sousa, A., & Sumuor, M. (2008). Majority-vote on directed Erdős-Rényi random graphs. *Physica A: Statistical Mechanics and its Applications, 387*(14), 3503-3510. doi:10.1016/j.physa.2008.01.120

Su, Z., Li, L., Peng, H., Kurths, J., Xiao, J., & Yang, Y. (2014). Robustness of Interrelated Traffic Networks to Cascading Failures. *Sci Rep, 4*(1). doi:10.1038/srep05413

Vasilev, V., Leguay, J., Paris, S., Maggi, L., & Debbah, M. (2018). Predicting QoE Factors with Machine Learning. *2018 IEEE International Conference on Communications (ICC)*, 1-6. doi:10.1109/ICC.2018.8422609

Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y., & Philip, S. Y. (2022). A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources. *IEEE Transactions on Big Data*, 415-436.

Xia, Y., Fan, J., & Hill, D. (2010). Cascading failure in Watts-Strogatz small-world networks. *Physica A: Statistical Mechanics and its Applications, 389*(6), 1281-1285. doi:10.1016/j.physa.2009.11.037

Zhou, C., Liu, Y., Liu, X., Liu, Z., & Gao, J. (2017). Scalable Graph Embedding for Asymmetric Proximity. *In Proceedings of the AAAI conference on artificial intelligence, 31*(1), 2943-2948.