Traffic Filtering(QoS) Dataset for SD-WAN

Vinayak Singh, Megha Dhar, Shivanshu Shrivastav and Asharani Chadchankar

May 31, 2022

Traffic Filtering(QoS) Dataset For SD-WAN

Vinayak Singh[1], Megha Dhar[2], Shivanshu Shrivastav[3], Asharani Chadcahnkar[4]

Student[1][2][3]

Professor[4]

NBN Sinhgad School of Engineering, Pune

Abstract

As the world is moving toward newer technologies and to meet the requirements of the same adapting toward different network topology. SD-WAN is such example of a network which solves many issues or limitations of a traditional TCP/IP network. As majority of workspace is moving towards SD-WAN, many new vulnerabilities are also being generated, and to protect the network and systems on these networks, in this paper we discuss and propose a dataset which would be helpful in training an intrusion detection system over SD-WAN which would also include the QoS Dataset for traditional TCP/IP network and hence on a hybrid network too. We generate this data over SD-WAN topology by attacking the host system present in the network, then analyze the generated data using CICFlowmeter which would give us the desired dataset for intrusion detection..

1 Introduction

In this project we have created an intrusion detection dataset by creating a virtual network environment and collecting the network traffic over while normal surfing as well as while attacking. The data packets are captured by wireshark and the pcap files generated by are then analysed by CICFlowmeter. This analyser gives us more then 80 features and no null or infinite values, hence best suited for our needs. After analysing the dataset generated is measured for accuracy by using various machine learning algorithms.

2 Literature Survey

This section reviews the existing publicly datasets generated from conventional networks. These datasets are widely used for intrusion detection in conventional networks, and they have been used for evaluating ML algorithms designed for anomaly detection approaches in SD-WAN networks.

KDD'99: one of the most well-known datasets which is used widely for intrusion systems evaluation. KDD'99 was derived from the DARPA packet traces. The dataset contains 41 traffic features which are classified into three groups: basic features, traffic features and content features. In addition, the dataset contains four attack categories, besides the normal data. The malicious traffic can be one of the following classes Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R), or probe attacks. One of the inherent problems in the KDD'99 dataset is the redundancy records, where the duplicated records in the training set reached to 78% and about 75% in the testing file. The high degree of duplication data prevents the detection techniques to give high accuracy for low attack categories like R2L and U2R. Thus, the detection systems are biased toward frequent records like DoS attacks.

NSL-KDD: NSL-KDD is the modified version of KDD'99 dataset. It was produced to solve some inherent problems in the KDD'99 dataset, such as duplicate records. NSL-KDD contains two subsets, training set and testing set. The distribution of attacks in the testing set is higher than the training one, with an additional 17 attacks that are not represented in the training set. Although many studies have employed KDD'99 and NSL-KDD in the domain of intrusion detection, both datasets are not realistic to represent the current network traffic since they were generated two decades ago and cannot reflect the current attack trends. Besides, the original DARPA dataset was generated using an outdated version of the TCP protocol. Using the old TCP version makes the header field ``IPv4 Type of Service (ToS)'' invalid according to modern standards. Besides the previous limitations of KDD'99 and NSLKDD datasets for IDS evaluation, they also have a large set of features that are not relevant to SD-WAN networks. For

example, some of the previous works [13] and [14] used six out of 41 features when deploying the NSLKDD dataset under the SD-WAN context. The both studies selected subset features that can be derived directly through the SD-WAN OpenFlow protocol. However, the performance of the classifier model indicates a low detection rate and a high false alarm as the used features are not being able to _nd the suspicious behaviour of malicious traffic. In addition, we can _nd most of the previous works in SD-WAN networks deployed KDD'99 and NSL-KDD datasets to identify DoS attacks only. This is because the other attack traffic like U2R and R2L are embedded in the packet's data, and the content features are required to identify these types of attacks. However, the content features are not directly accessible in the current OpenFlow protocol.

_ Kyoto dataset 2006C [15]: was collected from honeypot servers in Kyoto University. It contains the real network traffic in the period between (Nov. 2006 to Aug. 2009). Kyoto dataset comprises of 24 statistical features, 14 of them are shared with the KDD dataset. The background or normal traffic was created simultaneously with malicious traffic by deploying an additional server in the same honeypots network to produce a more realistic dataset. The imbalanced class distribution of the dataset is considered the main limitation of Kyoto 2006C since the traffic data was obtained from honeypot servers, and the majority of the traffics data are malicious. Besides, the attack types in the dataset are unknown. The shortcoming to identify the attack types gives a limited view to evaluate intrusion detection performance when using this dataset. Furthermore, the normal traffic in Kyoto 2006C covered only the mailing and DSN traces. In addition, the size of normal traffic in the dataset, i.e. between 3% and 4% of the whole dataset, does not reject the Internet traffic. Besides, normal and malicious traffics were created in two different environments causing to the dataset being unrealistic and uncorrelated [16]. Although the Kyoto 2006C dataset was built on real traffic data, it does not consider any information regarding the dataset attacks types. As a result, we can find difficulties in evaluating the impact of these attacks on the SD-WAN network services.

_ ISCX2012 [17]: The authors used two profiles to generate data traffic based on a simulated network environment. The Alpha-profiles are used to create attack traffic and Beta-profiles for normal traffic generation. The dataset includes two main types of network attacks, DoS and brute force attacks with 20 collected packet features. However, the diversity of the DoS attacks in the data is slightly small and does not cover the vulnerabilities that can be happened in different OSI layers. Furthermore, the dataset includes only HTTP traffic, which does not reject modern traffics, where the majority of current Internet traces are based on HTTPS traffic [18]. Again, similar to KDD'99 and NSL-KDD datasets, the number of features that can be extracted from the OpenFlow protocol are not enough for machine learning evaluation.

_ CICIDS 2017 [18]: This dataset is the closest one to our study due to it covers a comprehensive range of attack scenarios that are not addressed in the previous datasets, besides it contains the same number of gathered _ow-based features. Although the CICIDS 2017 dataset is considered one of the recent datasets that attracts many researchers to develop and analyse their new models, it contains many problems and shortcomings as the following: (i) Firstly, the CICIDS 2017 dataset was released based on the foundation of ISCX2012, published in 2012. The difference between both datasets is the total number of extracted features. Where the CICIDS 2017 dataset contains more than 80 _ow-based features compared to 20 packet features in ISCX2012. In addition, the HTTPS Beta profile was added to the CICIDS 2017 dataset to keep the adoption of HTTPS growth on the web. (ii) Secondly, normal traffic behaviour was generated based on profile scripts. However, applying the concept of profiling could be problematic due to their innate complexity [19]. Furthermore, Panigrahi et al. (2018) highlighted some problems and shortcomings in CICIDS 2017 data [20]. The dataset has 288602 missing class labels and 203 missing information instances. In addition, the size of the CICIDS 2017 dataset is extremely huge and contains many redundant records that seem to be irreverent for any IDS training.

_ CSE-CIC-IDS2018 [21]: The dataset is the result of a collaborative project between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC). Similar to CICIDS 2017 but instead, it was implemented on AWS (Amazon Web Services) computing platform. The notion of profiles is used to generate the dataset in a systematic manner. Where this dataset has two general classes of profiles, B-profiles is used to generate the normal traffic, and M-Profiles is used for attack scenarios. The dataset covers the same attack scenarios as in CICIDS 2017 dataset. However, the dataset suffers from the same inherent problems of CICIDS 2017, and also the use of synthetic traffic. In addition to datasets described above, many data repositories have been published to cover various security domains, such as botnets Malware Port scans , etc. While the structure and the type of those repositories are different, we exclude them from our comparison.

3 Implementation

This section presents our approach to generate the SD-WAN network traffic data by using different attack scenarios. The centralized view of the SD-WAN network and separation of the data plane from the control plane creates a new opportunity for the attacker to carry out various types of attacks compared to the conventional network. The nature of these attacks in SD-WAN is different from those commonly affecting the conventional network. For example, the attacker can generate new malicious traffic to attack the SD-WAN controller or even the communication links between the SD-WAN controller and OpenFlow switches.

Furthermore, compromised users can be employed to start a new attack after the traffic flow is established. Besides, the SD-WAN applications can have different vulnerabilities such as buffer overflow, command injection, SQL injection, etc. These vulnerabilities can create attack opportunities, and help the attacker to bypass the authentication mechanism, gain access to the controller through installing a malicious script. If the attacker successfully gains access to the controller, he can start new attacks such as flow rules manipulation, launching DoS attack, and eavesdropping on the data/control traffic. DoS attacks: Is one of the most common attacks inside

the SD-WAN architecture. It does not only damage the victim machine but can also overwhelm the SD-WAN controller resource in a short time. Besides, the SD-WAN controller is the brain of the SD-WAN network, and in the case of DoS attacks, the whole system becomes unavailable for legitimate users. It turns the entire network into a `body with no brain'. DoS attack can flood the victim machine with a huge number of spoofed packets that have no matched rules inside flow tables switches

Thus, the OpenFlow switch will send these flows to the SD-WAN controller in the form of packet-In message for further processing. When packet-In message rates are increased up to a certain limit, SD-WAN controller resources can be overwhelmed by many unprocessed packets. There are two main types of DoS attacks as the following:

• Network DoS attacks: The main objective of these attacks are to overwhelm the benign users by flooding the network bandwidth or victim machine by a large number of spoofed packets. The attacker often uses different protocols like UDP, TCP, or ICMP. DoS attacks can also disturb the SD-WAN controller or its channels due to the significant number of forwarded packets to the controller.

• Application DoS attacks: Even though these attacks do not require high bandwidth, however, it can cause serious damage to the target server and consume its resources in a short time. It mainly targets the top application layer or services such as HTTP. The application layer attack is not easy to detect since the intruder is connected to the victim server in an authorized manner.

Fig.8: dataset attack classes generated in virtual environment

7.Dataset Generation

We divided the dataset into three groups based on the traffic types and the target machines. The rst group includes normal traffic only. The second group contains the attack traffics that target Mealsplotable-2 server. In the last group, attacks on the OVS machine are considered. The Tcpdump tool is used to capture the traffic traces for each category at the target machine and the SD-WAN controller interface. In addition, the CICFlowMeter tool is used to extract the flow features for this dataset. The reason we decided to use the CICFlowMeter in our work despite many available tools in literature such as Argus 1 and Bro-IDS 2 is the fact that none of these tools exclusively consider the time-based features. However, different applications have different time constraints. As a result, it is more important to calculate the statistical time-related features for the ow traffics. The CICFlowMeter was generated by the Canadian Institute of Cybersecurity team and has been written in Java to create network flow traffics from the PCAP le. The generated flows are calculated in Bidirectional, where the first packet in the flow determines the flow direction (forward or backward).

The output of the CICFlowMeter is more than 80 statistical features in CSV file format such as Protocol, Duration, Number of bytes, Number of packets, etc. The list of extracted features and details are available in the appendix. We collected more than 80 features with 56 categories from our experiments. For simplicity, we divided the entire features into eight groups as the following:

- Network identifiers attributes: these features contain the common information that used to define the source and destination flow. For example, IP address, Port number, protocol type.

- Packet-based attributes: these features hold the information related to the packets such as the total number of packets in a forward and backward direction.

- Bytes-based attributes: these features hold the information related to the bytes i.e. total number bytes in the forward and backward direction.

- Interarrival time attributes: these features show the information related to the interarrival time in both forward and backward directions.

1. http://qosient.com/argus/index.shtml.

2. https://www.bro.org/index.html.

- Flow timers attributes: these features hold the information related to the time of each flow i.e. active and inactive. Flag attributes: these features hold the information related to the flags like SYN Flag, RST Flag, Push flag, etc.

• Flow descriptors attributes: these features contain the traffic flow information (eg., the number of packets and bytes in both forward and backward direction).

• Sub flow descriptors attributes: these features show the information related to sub flows, such as the number

of packet and bytes in forwarding and backward directions. For labeling processing, we use some features information such as Source IP and Destination IP. The total number of dataset instances are 343,939 for normal and attack traffic. Where the normal data brings a total of 68424, and attack traffic contains 275,515 instances. Table 5 represents the attack classes for each group with its total size. Furthermore, the name of PCAP files under each attack group is chosen based on the target protocol layer or the tools that are used to create each file

4 Conclusion and Discussions

This paper investigated the challenging problem related to the dataset availability in the SD-WAN environment. We proposed a new SD-WAN dataset to solve some of the inherent problems in legacy datasets. We considered different attack scenarios that represent the real-world scenarios and discussed the impact of the generated attacks on the different SD-WAN elements. We can observe that the SD-WAN can also be affected with the popular network attacks. However, the SD-WAN network is more sensitive to malicious traffic than the conventional environments. In the conventional network, any attacks can only affect the portion of the network almost for the same vendor without bringing down the entire network. However, in the SD-WAN environment, the compromised switches or end users can flood the SD-WAN controller, causing damage for the whole network. In the near future, we will extend this work and create a more intrinsic dataset generated from large-scale networks. Moreover, we will consider new attack categories for the best representative of existing real-world networks.

References

[1] M. S. Elsayed, N. -A. Le-Khac and A. D. Jurcut, "This: A Novel SD-WAN QoS Dataset," in IEEE Access, vol. 8, pp. 165263165284,2020,doi:10.1109/ACCESS.2020.3022633, doi:10.1126/science.1065467.

[2] O. Salman, I. H. Elhajj, A. Chehab and A. Kayssi, "QoS guarantee over hybrid SD-WAN/non-SD-WAN networks," 2017 8th International Conference on the Network of the Future (NOF), London, 2017, pp. 141-143, doi:

10.1109/NOF.2017.8251237

[3] R. Amin, M. Reisslein and N. Shah, "Hybrid SD-WAN Networks: A Survey of Existing Approaches," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 3259-3306, Fourthquarter 2018, doi: 10.1109/COMST.2018.2837161.

[4] A. Prakash and R. Priyadarshini, "An intelligent software defined network controller for preventing distributed denial of service attack", Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT), pp. 585-589, Apr. 2018

[5] D. Li, C. Yu, Q. Zhou and J. Yu, "Using SVM to detect DDoS attack in SD-WAN network", IOP Conf. Ser. Mater. Sci. Eng., vol. 466, Dec. 2018

[6] M. Banton, N. Shone, W. Hurst and Q. Shi, "Intrusion Detection Using Extremely Limited Data Based on SD-WAN," 2020 IEEE 10th International Conference on Intelligent Systems (IS), Varna, Bulgaria, 2020, pp. 304309, doi: 10.1109/IS48319.2020.9199950

[7] Habibi Lashkari, Arash. (2018). CICFlowmeter-V4.0 (formerly known as ISCXFlowMeter) is a network traffic Bi-flow generator and analyser for anomaly detection. https://github.com/ISCX/CICFlowMeter.

10.13140/RG.2.2.13827.20003.

[8] S. P. Bendale and J. Rajesh Prasad, "Security Threats and Challenges in Future Mobile Wireless Networks," 2018 IEEE

[9] Global Conference on Wireless Computing and Networking (GCWCN), Lonavala, India, 2018, pp. 146-150, doi: 10.1109/GCWCN.2018.8668635.

[10] Shailesh Pramod Bendale, Jayashree Rajesh Prasad. (2020). Security Challenges to provide Intelligence in SD-WAN with the help of Machine Learning or Deep Learning. International Journal of Advanced Science and

Technology, 29(05), 356 - 363. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/8983

[11] Chinmay Dharmadhikari, Salil Kulkarni, Swarali Temkar, Shailesh Bendale , Comparative Analysis of DDoS Mitigation Algorithms in SD-WAN , International Journal of Future Generation Communication and NetworkingVol.13,No.2s,(2020),pp.17001707http://www.sersc.org/journals/index.php/IJFGCN/article/ view/29228/16286