GS-EECS6414M Project Proposal

Jia Ying Ou*
caryou@my.yorku.ca
York University
Toronto, Ontario

Yunge Hao*
hyggs@my.yorku.ca
York University
Toronto, Ontario

ACM Reference Format:
Jia Ying Ou and Yunge Hao. 2020. GS-EECS6414M Project Proposal. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION AND DOMAIN DESCRIPTION

The Distributed Denial of Service (DDoS) attack is a well-known threat to network security. The DDoS attack is designed to prevent legitimate users from accessing network resources or a computer system. The first major DDoS kept Yahoo.com off the internet for about 2 hours, cost a potential loss of $500,000 [2] in the year 2000. According to the 2018 IDG report [3], there are 86% of people report experiencing one or more DDoS attacks and 70% are highly likely to consider changing their current solution to a higher effective solution. Therefore, analyzing complex network data to detect a DDoS attack is essential.

Many studies that have been done in detecting a DDoS attack. Our idea for this project comes from two primary sources. First, Sharafaldin et al. [11] addressed several shortcomings that appeared in the dataset of previous studies, such as the lack of finding a comprehensive dataset for detection model evaluation. A new dataset called CICDDoS2019 [11] is generated to remedy the limitation of previous datasets. It is generated by a testbed that simulates real-world network attacks. There are two parts to the dataset. One part is the training data, and one is the testing data. In the training data, it consists of benign traffics and 11 types of attacks, including NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, SYN and TFTP. In the testing data, it consists of benign traffics and 7 types of attacks, including PortMap, NetBIOS, LDAP, MSSQL, UDP, UDP-Lag and SYN. There are 80 features extracted from the dataset using a tool called CICFlowMeter [4], such as source IP address, destination IP address, etc. They built their model using machine learning algorithms includes ID3, RF, Naïve Bayes and Logistic Regression for pattern capturing. However, the results are not robust enough.

Hence, our project will be using a completely different approach to analyze the dataset and detect DDoS attacks. This approach is from the second source [7]. Our project aims to analyze the differences between Traffic dispersion graphs (TDG) in time series to detect malicious activities and using the VF2 isomorphism algorithms to identify attack patterns in anomalous traffic.

The potential applications are in the domain of network security. One of the applications can be a server with a built-in attacker-detecting algorithm that helps it recognize certain requests to avoid overloading. Another application is that visualization identifies and characterizes problems to effectively increase operators' situation awareness, letting them detect and respond to malicious activities in a quicker manner. And the DDoS attack classifier shortens the time for the network operators to discover the pattern and serves as an aid when the operator is under the decision-making process for faster mitigation.

2 METHODOLOGY

The data analytic method that we will be using belongs to graph mining. It has advantages over statistical and machine learning in not being subject to parameters adjusting. In the case of detecting network anomalies, we adapt graph theory to reduce the chance of producing false alarms [7].

Our project consists of two parts. The first part is anomaly detection, and the second part is anomaly identification. The methods are based on the paper "Traffic Dispersion Graph Based Anomaly Detection" by Le et al [7].

For the anomaly detection part, we will be using a traffic dispersion graph (TDG) to model the network traffic from our dataset. There are 80 features in our dataset, we will be extracting 5 of them, and use it as the standard 5-tuple srcIP, dstIP, srcPort, dstPort and protocol for analysis. The below set of graph metrics will be used to analyze the TDGs:

(1) Static Metrics [6]
(2) Node degree
(3) Vino, Vin, Vout [6]
(4) Maximum degree(Kmax)
(5) The entropy of the degree distribution
(6) Dynamic Metrics
(7) Graph edit distance [5] [9]
(8) dK-2 distance metric [10] [8]

A TDG visualization for the attack pattern will be provided.

For the anomaly identification part, the problem is a multi-label classification problem since its task is to find out the cause for the anomaly. More specifically, it not only distinguishes malicious attacks from benign network traffic but classifies different types of DDoS attacks, such as LDAP, Syn, UDP, DNS, etc. To identify the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

attack, we first obtain the attack structure pattern that TDGs were generated. Then, we use the graph matching method for identification. In our approach, we use the VF2 algorithm. VF2 algorithm can be used for both graphs and sub-graph isomorphism to find out if one object is part of another object. We will be applying VF2 to identify attack patterns in abnormal TDGs for a faster approach since attack patterns are located in the abnormal traffic.

3 EVALUATION

The evaluation section consists of two parts. The following metrics will be used for the first part to make comparison of the result when utilizing different graph metrics.

- Maximum degree(Kmax)
- dK-2 distance
- the number of packets
- measured over time of trace of the dataset

In addition, the classification results from VF2 will be presented in a multi-label confusion matrix as a heatmap.

We will be validating our algorithm in the second part by using the testing dataset from the CICDDoS2019 dataset. The dataset CICDDoS2019 is generated in one day within 2 to 7 hours of duration [11], but our method and analysis allow it to be scaled for longer time interval of data. Since the testing dataset is smaller than the dataset that we use in the methodology part, thus, we will be scaling down during the evaluation process.

The three evaluation metrics are precision, recall and F-Measure.

- Precision, or positive predictive value, is the correct classification count of flows (TP) divided by total classifications (TP+FP).

$$Pr = \frac{TP}{(TP + FP)}$$

- Recall or sensitivity is the correct classification count (TP) divided by all generated flows (TP+FN).

$$Rc = \frac{TP}{(TP + FN)}$$

- F-Measure (F1) is a harmonic combination of precision and recall, ranging from 0 to 1, with 1 being the most desirable.

$$F1 = \frac{2}{(\frac{1}{Pr} + \frac{1}{Rc})}$$

Other datasets that can be used for evaluation. They are "CAIDA DDoS Attack 2007" dataset, and the DARPA dataset [1] mentioned as previous research. Also, from [7], there are traffic traces from POSTECH in July 2009 and CAIDA DDoS trace in 2007, which are the exact dataset used by [7].

REFERENCES

[1] 2000. DARPA Intrusion Detection Scenario Specific Datasets | MIT Lincoln Laboratory. https://www.ll.mit.edu/r-d/datasets/2000-darpa-intrusion-detection-scenario-specific-datasets

[2] 2000. Yahoo on Trail of Site Hackers | WIRED. https://www.wired.com/2000/02/yahoo-on-trail-of-site-hackers/

[3] 2018. IDG DDoS REPORT. https://www.a10networks.com/wp-content/uploads/a10-eb-14116-en-evolving-strategies-for-handling-todays-complex-costly-threat.pdf?mkt{_}tok=eyJpIjoiTkdZeU5UQXlZbUkxTmpGbCIsInQiOiJFM3JWNkx6UndMa2FkcXB6N\k5NTSt3M0RNYTIxWEdobUwrY1NYd3FpNFVLK2UxMEtmQ1plWERQOE8wV2tn\YStYTU5EQXVvYUJsa2lneGxcL0VobDY1Z1g0bGdcL2tkOHlMaHpwcXZWWlh\jajUzVEZlN1VvRGpRUmlzS21YaG9PeTlpIn0{%}3D

[4] Lashkari Arash H., Zang Yongcan, Owhuo Gift, Mamun Mohammad S.I., and Gerard D. Gil. 2015. Network Traffic Flow analyzer. http://netflowmeter.ca/netflowmeter.html

[5] Marios Iliofotou, Michalis Faloutsos, and Michael Mitzenmacher. 2009. Exploiting dynamicity in graph-based traffic analysis: techniques and applications. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*. 241–252.

[6] Marios Iliofotou, Prashanth Pappu, Michalis Faloutsos, Michael Mitzenmacher, Sumeet Singh, and George Varghese. 2007. Network monitoring using traffic dispersion graphs (tdgs). In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 315–320.

[7] Do Le, Taeyeol Jeong, H. Roman, and James Hong. 2011. Traffic dispersion graph based anomaly detection. *ACM International Conference Proceeding Series*, 36–41. https://doi.org/10.1145/2069216.2069227

[8] Priya Mahadevan, Calvin Hubble, Dmitri Krioukov, Bradley Huffaker, and Amin Vahdat. 2007. Orbis: rescaling degree correlations to generate annotated internet topologies. In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*. 325–336.

[9] Panagiotis Papadimitriou, Ali Dasdan, and Hector Garcia-Molina. 2010. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications* 1, 1 (2010), 19–30.

[10] Alessandra Sala, Lili Cao, Christo Wilson, Robert Zablit, Haitao Zheng, and Ben Y Zhao. 2010. Measurement-calibrated graph models for social network experiments. In *Proceedings of the 19th international conference on World wide web*. 861–870.

[11] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali Ghorbani. 2019. Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy. 1–8. https://doi.org/10.1109/CCST.2019.8888419