Vertical Scanning Behavior Analysis of High-Frequency Superpoints

Wenxian Guo
School of Cyber Science and Engineering
Southeast University
Nanjing, China
wxguo@njnet.edu.cn

Haiqing Yu
School of Cyber Science and Engineering
Southeast University
Nanjing, China
hqyu@njnet.edu.cn

Wei Ding
School of Cyber Science and Engineering
Southeast University
Nanjing, China
wding@njnet.edu.cn

***Abstract*—Access superpoint is a host that communicates with a large number of peers at the same time in the network, occupying a large number of network communication resources. Under the background that access superpoint detection algorithms have been developed relatively mature, the anomaly detection research based on this is the direction worth exploring at present. In terms of time, access superpoints can be divided into high-frequency, medium-frequency and low-frequency superpoints. Among them, high-frequency superpoints often contain important data resources and are the first choice for hackers to attack, while vertical scanning is a common pre-invasion method for attackers. Therefore, detecting and analyzing the vertical scanning behavior of high-frequency superpoints plays an important role in the protection of high-frequency superpoints. In this paper, a time-frequency attribute is defined for the detected access superpoints and a time-frequency classification algorithm based on sliding window is proposed. The experimental results show that the algorithm has a high accuracy of 98.26% in a high-speed network environment. The vertical scanning behavior was screened based on the rules. And XGBoost algorithm was used to generate a classifier that can distinguish the abnormal behaviors of high frequency superpoints caused by vertical scanning. The classifier can identify the abnormal behaviors of high frequency superpoints caused by vertical scanning with an accuracy of 93.19%.**

Keywords—internet traffic measurement, superpoints, time-frequency classification, vertical scanning, machine learning

I. INTRODUCTION

Access superpoints are hosts that interact with a large number of peer hosts on a network. They generally play an important role in a network, such as servers, agents, scanners, hosts attacked by DDoS attacks, and worm propagation sources. Although the number is small, they occupy a large number of network communication resources. It is very important for network security and network management to detect the access point effectively and monitor the traffic behavior in real time.

The main problem faced by access superpoint detection in high-speed network environment is the contradiction between limited computing storage capacity and increasing massive data. The effective integration of Sketch technology and the continuous upgrade of sketch data structure are helpful to solve the problem [1]-[4]. With the support of Sketch theory and GPU platform, the detection problem of access superpoint has been basically solved. However, the work of network security management based on access superpoint detection is less. Obviously, in the background of access superpoint detection algorithm has been relatively mature, how to apply it to the actual Internet security management is worth studying at present.

In terms of time, each access superpoint has different active heat and duration, which directly reflect the behavioral patterns of superpoints on the time axis. Access superpoints can be divided into high-frequency, medium-frequency and low-frequency superpoints based on the activity degree on the time axis. High-frequency superpoints usually play specific roles or carry specific applications, such as Web servers, NAT servers, P2P applications, and DNS servers. The formation of such access superpoints is caused by normal network behaviors, and the traffic behavior pattern is persistent and stable. At the same time, this kind of superpoint is often the first choice of hacker attack because of its important data resources.

Vertical scanning is a kind of pre-invasion method commonly used by attackers. Attackers can obtain the basic information of target hosts through vertical scanning. Therefore, detecting and analyzing the vertical scanning behavior of high-frequency access superpoints plays an important role in the protection of high frequency access superpoints.

In this paper, the time-frequency attribute is defined for the access superpoints, and a time-frequency classification algorithm based on sliding window is proposed. The algorithm is tested in the actual high-speed network environment, and the vertical scanning behavior of the high-frequency access superpoints is detected and analyzed on the basis of the algorithm.

II. RELATEDWORKS

A. Superpoints

Access superpoint refers to a network node that communicates with other hosts whose number is much larger than the average in a certain period of time, and is a kind of core host in the network [5].

Access superpoint detection is a classic research area, so far, the research work has achieved a lot of results. In [6], Wang et al. proposed DCDS (Double Connection Degree Sketch) algorithm in 2011. In [7], Li et al. proposed an ESC detection scheme including bitmap and maximum likelihood estimation in 2013. In [8], Liu et al proposed a data stream valuation algorithm based on VBF (Vector Bloom Filter) structure in 2015. In [3], Zhou Aiping et al proposed a parallel data stream algorithm based on sketch structure in 2016. In [4], Xu Jie proposed SRLA (Sliding Rough and Linear Algorithm) algorithm in 2019. This algorithm is designed for sliding window. It can output the access superpoints in seconds at the boundary of high-speed network. Compared with the traditional discrete window algorithm, it can output rich access superpoints' information more smoothly, which makes it possible to study the behavior of access superpoints from the perspective of time and frequency.

B. Vertical Scanning Detection

Vertical scanning is an extremely important means of pre-attack detection. Through vertical scanning, information such as open ports and running services of the target host can be obtained [9].

Threshold-based scanning detection method is the most widely used.This method was first applied to Network Security Monitor(NSM) intrusion detection system in [10]. In [11], Jung et al. proposed TRW algorithm, a test algorithm based on sequential hypothesis in 2004. In [12], Li Gang proposed a scanning behavior detection algorithm based on port matching and scanning flow characteristics in 2016. Rule-based detection technology, that is, by setting rules in the knowledge base in advance, analyzing the network data traffic, and extracting the network behavior of matching data. In [13], Mahoney et al. proposed PHAD(Packet Header Anomaly Detection) in 2002, which detects scanning events by giving an exception score to the value combination rule of all fields of packet header in ethernet. In [14], Kim et al. proposed an abnormal flow control framework that can be used to detect vertical scanning behavior using fuzzy rules in 2008. All the above researches have laid the foundation for the development of vertical scanning behavior detection. In this paper, the vertical scanning behaviors of access superpoints were screened based on rules, and combined with the analysis of the behavior pattern measure of high-frequency superpoints, the XGBoost (Extreme Gradient Boosting) algorithm was used to realize the vertical scanning classifier.

C. Access Superpoints and Anomaly Detection

Access superpoint anomaly detection is a branch of abnormal network traffic detection, and most of its detection methods are based on the research work of abnormal network traffic detection. Abnormal network traffic detection is usually aimed at network attacks related to network traffic activity rules such as DDoS attacks and malicious scanning, and its common methods include detection methods based on statistical analysis, classification and clustering [15].

The anomaly detection of access superpoint is of great significance to network security and network management. In [16], Jiang et al. analyzed the network traffic accessing the overstay based on the sliding window mechanism, and measured the similarity of the traffic according to the distribution distance of the traffic on the time axis in 2015. In [17], Vladimir et al. used the cross-correlation function to calculate the correlation of the time series of the bytes size of the request and response to detect the exceptions of the network server in 2016. The above two studies are based on static data sets or simulation environment, due to the complexity of the actual network environment, it is difficult to achieve the expected results in the measured environment. Therefore, the research of access superpoint anomaly detection under the actual network environment has very large exploration space.

From the view of the cause of access superpoints, it can be simply divided into normal and abnormal. However, this statement only gives the distinction conceptually, no specific attribute is determined and quantified. In this paper, the time-frequency attributes of access superpoints are defined and quantified from the time dimension. According to different time-frequency attributes, access superpoints are classified systematically. Based on different categories of access superpoints, its behavioral characteristics can be better studied. For example, high-frequency superpoints have stability and persistence, and most of them are hot servers, while low-frequency superpoints have sudden onset, often caused by specific network exception events.

III. Time-Frequency Classification Algorithm of Access Superpoint Based on Sliding Window

The occurrence of some abnormal network events may cause related hosts to become access superpoints, which tend to show low active heat. While access superpoints formed by specific applications,such as some servers, have high active heat. The two types of access superpoint anomaly detection focus on distinct differences according to their different time-frequency attributes. Therefore, time-frequency classification is the basis of access superpoint anomaly detection.

A. Definition

In the last consecutive M time windows, if a specific host becomes the access superpoint in N of the windows, then the access superpoint time frequency Freq(IP)=N/M, and IP is the address of the specific host. The time-frequency attribute is used to quantify the activity of the access superpoint.

B. Algorithm Description

In the high-speed network environment, the time-frequency classification of all detected access superpoints is carried out. If the accurate statistical method is used, it will occupy a large memory space due to the excessive number of access superpoints or the large number of historical time Windows observed. The exponential histogram algorithm [18] of real-time statistics of data stream can solve this problem.In this paper, a

time-frequency classification algorithm based on sliding window is proposed.

The main idea of the algorithm is to construct a histogram containing some buckets for each access superpoint, and each bucket records its arrival timestamp and frequency information. For the input of the same access superpoint, first of all, determine whether a timestamp in the last bucket in the current window, if not then remove the bucket, then, to create a frequency of 1 new bucket containing the current timestamp. Traverse the bucket list, if there are more than $\frac{C}{2}$+2 buckets of the same frequency, merge the two buckets with the smallest timestamp. In other words, the new bucket whose frequency is the sum of the two buckets contains the timestamp closer to the current one, and the two buckets are deleted. Finally, the time frequency of the access superpoint can be calculated according to the formula as in (1), where Sum is the frequency sum of all buckets, and f(k) is the frequency of the bucket with the smallest current timestamp.

$$\mathrm{Freq} = \mathrm{Sum} - \frac{\mathrm{f(k)}}{2} \qquad (1)$$

The statistical error of frequency generated by this algorithm only comes from the estimation error of the last bucket, and the relative error of frequency obtained can be controlled within ε by using $(\frac{\mathrm{C}}{2}+1)(\log(\frac{2\mathrm{W}}{\mathrm{C}}+1)+1)$ buckets, where $\mathrm{C} = \left\lceil\frac{1}{\varepsilon}\right\rceil$. In [18], this conclusion has been confirmed by Mayur et al.

C. Algorithm Validation

In order to verify the accuracy of the above algorithm, this paper tests the algorithm for full traffic data on the 50G bandwidth CERNET Nanjing main node network boundary.

In practical applications, it is necessary to obtain the threshold to distinguish high-frequency, medium-frequency and low-frequency for time-frequency classification of access superpoints. Therefore, first, two thresholds, T1 and T2, need to be determined. The frequency not greater than T1 is the low-frequency superpoint, the frequency not less than T2 is the high-frequency superpoint, otherwise, it is the intermediate-frequency superpoint. Based on SRLA superpoint detection algorithm, access superpoints per second at the network boundary are obtained in real time. The access superpoint data of 7 days, was recorded in the unit of days. The experiment time was from July 29, 2021 to August 4, 2021. The cumulative distribution function (CDF) curve is obtained by counting the access superpoint data every day and calculating the frequency of each access superpoint.

Through observation, the access to the superpoint time-frequency CDF curve has two obvious inflection points, according to these two inflection points, this paper determines the value of the day's threshold, and the final threshold is generated based on the weekly mean value.After experiments, the thresholds are 0.00471 and 0.57643, respectively.

Next, the time-frequency classification algorithm is tested. The time window is set to 1 day, that is, 86400s. The access superpoint on August 5, 2021 is classified in time-frequency, and the access superpoint on this day and its time-frequency classification result determined by the algorithm are recorded.Based on the data of August 4, the actual time-frequency attribute of access superpoint on August 5 can be calculated by sliding window. In the experiment, data of 5 minutes in each hour were randomly selected for statistical analysis and compared with the results obtained by the algorithm. The precision rate, recall rate and overall accuracy of different categories of superpoints were obtained, and the mean value was taken to obtain the final results, as shown in Table 1. The accuracy of the algorithm can reach 98.26%.

TABLE I. ACCESS SUPERPOINT TIME-FREQUENCY CLASSIFICATION EVALUATION

Time-Frequency	Precision	Recall
High-Frequency	97.22%	97.47%
Medium-Frequency	98.80%	98.60%
Low-Frequency	95.36%	99.94%

IV. VERTICAL SCANNING IDENTIFICATION OF HIGH FREQUENCY ACCESS SUPERPOINTS

A. The Overall Train of Thought

In this paper, high-frequency access superpoints are mainly studied. The formation of such access superpoints is mainly due to normal network behavior, and the behavior pattern of its traffic is persistent and stable. The intervention of abnormal traffic will lead to a large deviation between the traffic behavior of high-frequency access to the superpoint and its normal traffic behavior mode.

Firstly, 100 IP addresses are selected. Then, the superpoint detection algorithm and time-frequency classification algorithm are applied to determine the superpoint and time-frequency classification of these IP addresses.

At the same time, the rule is used to determine the vertical scanning condition of these hosts. Combined with the actual needs, in view of the high-frequency superpoints, set up corresponding detection index and the rules are as follows:

1) The number of different ports that one peer IP address sends packets exceeds a.

2) The average number of packets received by each destination port does not exceed b.

3) Each destination port can receive a maximum of c packets.

4) The average number of bytes in a packet does not exceed d.

5) There are no more than e types of packets with different bytes.

6) The proportion of incoming unidirectional packets exceeds f.

If the above conditions are met, the high-frequency superpoint is judged to be subjected to vertical scanning, and the above a, b, c, d, e and f are taken as the index threshold.

Then, appropriate behavior measures are selected to model the high frequency superpoint behavior pattern. In this paper, we select flow statistical measures, including the number of packets arriving per second, average length of stream, the arrival time interval of packets, ratio of single packet flow, etc. And we choose various entropy values, including IP entropy and port entropy. On the basis of rules-based decision, the time granularity of the high-frequency superpoint subjected to vertical scanning is located, and the key measure of the anomaly is found according to the data of the upper and lower time granularity. If the entropy of the inbound destination port increases, the average number of incoming single packets decreases, the average length of incoming packets decreases, the proportion of incoming unidirectional flow increases, and the proportion of incoming single packet flow increases, the vertical scanning behavior is considered to affect the normal behavior mode of the high-frequency superpoint.

Finally, machine learning algorithm can be used to generate a high-frequency superpoints vertical scanning classifier to realize the determination of high-frequency superpoints vertical scanning anomaly.

B. *Experimental Results and Analysis*

The experiment was carried out on the network boundary of CERNET Nanjing master node with 50G bandwidth. SRLA superpoint detection algorithm based on GPU platform and time-frequency classification algorithm based on sliding window were applied. In order to get more data, with the high-frequency superpoints of stability and continuity, choose 100 IP addresses that have become high-frequency superpoints. Each time granularity is 5 minutes.Superpoint determination and time-frequency classification were carried out for these IP addresses. And 32 behavior measures were selected to describe the traffic behavior pattern of high-frequency superpoints.

The data on the time granularity of the high-frequency superpoint with vertical scanning behavior were combined with the upper and lower time granularity for average normalization, and the samples with abnormal high-frequency superpoint caused by vertical scanning behavior were obtained through manual analysis to constitute the data set.

In this paper, XGBoost algorithm is selected to generate the classifier. As an improvement of GBDT (Gradient Boosting Decision Tree), this algorithm has better learning effect and training speed, and is widely used in classification and regression problems. The abnormal sample data were divided according to the ratio of 7:3, with 70% as the training set and the remaining 30% as the test set. The anomaly classifier was constructed based on the training set, and the classification effect was evaluated through the test set. The above steps were repeated, and the classification accuracy of the classifier was up to 93.19%. Its ROC (Receiver operating characteristic curve) is shown in Fig.1. Meanwhile, after feature selection by the algorithm, the number of sample features is reduced from the initial 32 to 13, which indicates that this classifier has a good classification effect.

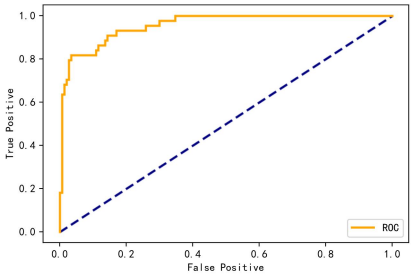

Fig. 1. ROC of Classification

V. CONCLUSION

In this paper, the vertical scanning behavior of high-frequency superpoint is analyzed, and the practice is carried out in the actual high speed network environment. In actual network environment, the time-frequency classification algorithm based on sliding window was tested, the experimental results show that the algorithm has high accuracy. At the same time, a classifier was generated to identify the abnormal behaviors of high-frequency superpoints caused by vertical scanning. The classifier has a good classification effect. In the future, we will analyze the intention of scanning hosts and filters out malicious host packets to further improve the protection performance against high-frequency superpoints.

REFERENCES

[1] Cheng Guang,Qiang Shi-Qing.Super point detection based on sampling and data streaming algorithms.Journal of Southeast University(English Edition),2009,25(02):224-227

[2] Zhang Qi.Improvement of the algorithm for detecting superpoints using VBF[Master Dissertation].Dalian Maritime University,Dalian,2017

[3] Zhou Ai-Ping,Cheng Guang,Guo Xiao-Jun,Liao Yi-Xin.Parallel data streaming method for detection of super points in high-speed networks.Journal of Software,2016,27(07):1841-1860

[4] Xu Jie.Research on access super points detection under high speed network [PhD Dissertation]. Southeast University, Nanjing, 2019

[5] Venkataraman S,Song D,Gibbons P B,et al.New streaming algorithms for fast detection of superspreaders//Proceedings of the Network and Distributed System Security Symposium.San Diego,USA,2005:149-166

[6] Wang P,Guan X,Tao Q,et al. A data streaming method for monitoring host connection degrees of high-speed links.IEEE Transactions on Information Forensics & Security,2011,6(3):1086-1098

[7] Li T,Chen S,Luo W,Zhang M,Qiao Y.Spreader classification based on optimal dynamic bit sharing.IEEE/ACM Transactions on Networking,2013,21(3):817-830

[8] Liu W, Qu W, Gong J, et al. Detection of superpoints using a vector bloom filter.IEEE Transactions on Information Forensics and Security, 2015,11(3):1-1

[9] Tang Xiao-Ming,Liang Jin-Hua,Jiang Jian-Chun,Wen Wei-Ping.Research about technology of port scan and port scan detect.Computer Engineering and Design,2002(09):15-17(in Chinese)

[10] Heberlein,L.Todd,et al.A network security monitor//Proceedings of the 1990 IEEE Computer Society Symposium on Research in security and privacy.Oakland,USA,1990:296

[11] Jung J,Paxson V,Berger A W,et al.Fast portscan detection using sequential hypothesis testing//Proceedings of the 2004 IEEE Symposium on Security and Privacy.Berkeley,USA,2004:211-225

[12] Li Gang.Research of scanning and drdos attack detection based on netflow[Master Dissertation].Southeast University,Nanjing,2016

[13] Mahoney M V,Chan P K.PHAD:Packet header anomaly detection for identifying hostile network traffic//Proceedings of the 2002 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.Edmonton,Canada,2002

[14] Kim J,Lee J H. A slow port scan attack detection mechanism based on fuzzy logic and a stepwise plolicy//Proceedings of the 4th IET International Conference on Intelligent Environments.Seattle, USA,2008:1-5

[15] Gong Jian,Yang Wang.Introduction of computer network security.The 3rd Edition.Nanjing:Southeast University Press,2020

[16] Jiang H,Chen S,Hu H,et al.Superpoint-based detection against ditributed denial of service(DDoS) flooding attacks//Proceedings of the 21st IEEE International Workshop on Local and Metropolitan Area Networks.Beijing,China,2015:1-6

[17] Eliseev V,Gurina A.Algorithms for network server anomaly behavior detection without traffic content inspection//Proceedings of the 9th International Conference on Security of Information and Networks.New Jersey,USA,2016:67-71

[18] Datar M,Gionis A,Indyk P,et al.Maintaining stream statistics over sliding windows.SIAM Journal on Computing,2002,31(06):1794-1813