Dynamic Group Behavior Analysis and Its Application in Network Abnormal Behavior Detection

Yan Tong[1(✉)], Jian Zhang[1], Wei Chen[1], Mingdi Xu[2], and Tao Qin[3]

[1] System Research Department, Wuhan Digital Engineering Institute, 718 Luoyu Road, Hongshan District, Wuhan, China
tongyan.cherish@l39.com, richardxx@l26.com, 772382203@qq.com

[2] System Software Department, Wuhan Digital Engineering Institute, 718 Luoyu Road, Hongshan District, Wuhan, China
mingdixu@l63.com

[3] School of Electronic and Information Engineering, Xi'an Jiaotong University, 28 Xianning West Road, Xi'an, Shaanxi, China
qin.tao@mail.xjtu.edu.cn

Abstract. Focus on the difficulty of large-scale network traffic monitoring and analysis, this paper proposed the concepts of Group Behavior Flow model to aggregate traffic packets and perform abnormal behavior detection. Based on the flow model the pivotal traffic metrics can be extracted while the number of flow records are reduced significantly. Secondly, we employ the graph model to capture the traffic feature distribution between different group users. And optical flow analysis methods are proposed to extract the dynamic behavior changing features between different groups and achieve the goal of abnormal behavior detection. The experimental results based on actual traffic traces show that the methods proposed in this paper can capture the traffic features effectually in the current 10 Gbps network environment, and achieve the goal of abnormal behavior detection and abnormal source location, which is very important for traffic management.

Keywords: Group user model · Dynamic behavior · Optical flow analysis Abnormal detection

1 Introduction

It is an important task that how to capture and analysis abnormal network traffic in order to maintain network under control, but it is more and more difficult to capture and analysis the abnormal traffic with the increase of number of users and bandwidth of backbone networks [1]. How to deal with the massive traffic data and detect the abnormal behavior effectively is an urgent problem to be solved.

Analysis the statistical traffic characteristics (such as the total number of packets per unit time, bytes, etc.) can not satisfied the need for traffic monitoring in large scale networks, many abnormal behavior detection methods based on suddenly change are ineffective [2–4]. In order to overcome this difficulty, CISCO proposed the conception

B. Li et al. (Eds.): ChinaCom 2017, LNICST 237, pp. 292–301, 2018.
https://doi.org/10.1007/978-3-319-78139-6_30

of Netflow [5], based on the Netflow framework many detailed features can be extracted from the raw traffic, but with the rapid growth of network traffic, even a medium-size enterprise LAN can generate about millions Netflow records per minute, which bring great difficulties for mining network traffic characteristics effectively. To solve those challenges, Kim and Reddy [6] proposed a method to describe network traffic characteristics based on image, which can monitor network traffic characteristics effectively, abnormal detection methods suit for large scale networks are also proposed. In order to further reduce the difficulty to monitor large-scale network traffic, Lakhina proposed the concept of ODFlow in [7], which divides the network into different domains with AS (Autonomous System), all packets through the border routers will be aggregated for different ODFlow records according to the source AS and destination AS. However, this method is only suitable for detecting abnormal behavior in autonomous domains and can't be applied to large-scale enterprise LAN.

To solve the above problems, this paper proposed the concepts of Group Behavior Flow Model to describe the network traffic characteristics, group users can be defined by different network address-prefix, such as using 24-bit address-prefix to aggregate network IP address, which can generate different scale network group users. The packets which have the same source group and destination group are aggregated into a group behavior flow within a certain statistical interval. The distribution of network traffic on different logical links consisted by different groups can be described by a graph, and different elements of the graph describe the characteristics of communication relationship between different group behaviors. The proposed methods can detect abnormal behaviors based on the dynamic change of the feature distributions between adjacent time. The experimental results based on actual traffic traces show that the methods proposed in this paper can capture the traffic features effectually in the current 10 Gbps network environment, and achieve the goal of abnormal behavior detection and abnormal source location.

2 The Concept of Group Behavior

The results of the literature [8] show that more than 98% of IP addresses are no more than 1 km in physical distance for the same C network segment. The survey results about the IP address distribution based on the campus network of Xi'an Jiao Tong University show that the IP address in the same C network segment is distributed in the same dormitory building, that is to say the physical distance is very close. Also the students from the same school are often in the same dormitory, which shows the network users in the same C network segment have the same network requirements, which also provide a reasonable explanation for splitting users in to the group flow model according to the C network segment.

Figure 1 presents the definition of group behavior flow, according to the monitoring location the network are divided into two parts, the internal monitoring network and external network, the internal monitoring network is the network need to be monitored, which is the campus network of Xi'an Jiao Tong University, the external network is the external Internet which composed by the monitoring network communicate with. The network topology can be described by the logical links formed between groups, then

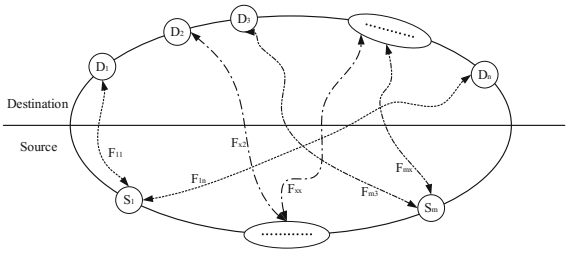

Fig. 1. Group behavior flow

the monitoring of network abnormal behaviors can be implemented according to the network traffic distribution on the logical topology. The paper only focuses on the communication between internal network and external network, and the change of the interaction behavior between the internal users and external network.

The variable d represents the destination group, s shows the source group, F means the network flow between groups, and such as F_{11} indicates the network flow from source group s_1 to destination d_1. We can describe the network logical topology using the below matrix according to the definition of group behavior flow.

Region	d_1	d_2	$d_{..}$	d_n
s_1	F_{11}	F_{12}	$F_{1..}$	F_{1n}
s_2	F_{21}	F_{22}	$F_{2..}$	F_{2n}
$s..$	..	..	..	..
s_m	F_{m1}	F_{m2}	F_{m3}	F_{mn}

The above matrix uniquely defines the interaction behavior between internal and external network. As time goes on, the change of the element of the above matrix can reflect the changes of the network traffic distribution. The paper will use this change to achieve network traffic monitoring.

3 Algorithm Framework of Network Traffic Feature Extraction

According to the above definition of group behavior flow, the processing framework of this paper is described as Fig. 2.

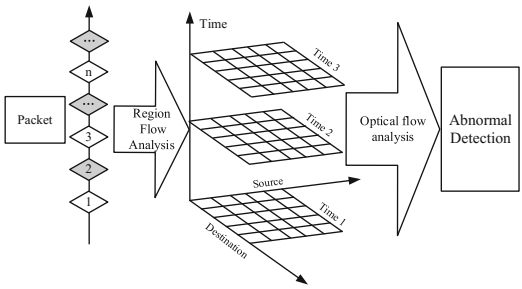

Fig. 2. The algorithm framework based on group behavior flow

Figure 2 is the algorithm framework of network traffic monitoring based on group behavior flow. First of all, according to the definition of network group flow, we regard the network packet sequence as group behavior flow, and then form the distribution matrix which can describe network traffic characteristics. As time goes on, it can form the traffic characteristics graph for different time, the behavior monitoring of dynamic network and extraction of behavior features can be implemented using the method of optical flow analysis which can analyze the dynamic change characteristics of network traffic among the groups with the change of time.

4 Description of Group Network Behavior Feature

4.1 Description of Group Network Behavior

Network traffic is generated by users, and this paper will describe network behavior by the dynamic characteristics of network traffic statistics. Supposing the sampling time is the variable t, $Byte_{ij}(t)$ means the total amount of data transmitted between the source region S_i and the destination region d_j, and the definition of data distribution function as shown as formula (1)

$$\begin{cases} p(s_i, d_j, t) = \frac{Byte_{ij}(t)}{\sum_{i,j} Byte_{ij}(t)} \\ \Delta p(s_i, d_j) = \Delta p(s_i, d_j, t) - \Delta p(s_i, d_j, t-1) \end{cases} \tag{1}$$

The distribution function is uniquely defined the distribution of the logical links which formed by the communication relationship among different groups at time t, the distribution can reflect the characteristics of users behavior in a certain time point, when the network is running normally, the distribution has certain stability, namely the changes of network data in the logical link are very small, or is approximately stable, this paper summarizes the change of distribution as follows:

(1) The stability of group behavior: the situation can describe the most common habitual behavior in the network, namely the network user behaviors are relatively stable in a short time interval, and the behavior changes gently, and the case is described as follows:

$$\begin{cases} p(s_i, d_j, t-1) \neq 0 \\ p(s_i, d_j, t) \neq 0 \\ \Delta p(s_i, d_j) \approx 0 \end{cases} \tag{2}$$

(2) The change of network group behavior, the abrupt change of network behavior is the main information source of network abnormal behavior monitoring,it can be described as formula (3).

$$\begin{cases} p(s_i, d_j, t-1) \neq 0 \\ p(s_i, d_j, t) \neq 0 \\ \Delta p(s_i, d_j) \neq 0 \end{cases} \tag{3}$$

(3) The disappearance of network group behavior,which represents the disappearance of network behavior among a particular group, and it describes as formula (4).

$$\begin{cases} p(s_i, d_j, t-1) \neq 0 \\ p(s_i, d_j, t) = 0 \end{cases} \tag{4}$$

(4) The birth of network group behavior, which represents the birth of network behaviors among specific groups, and it describes as formula (5).

$$\begin{cases} p(s_i, d_j, t-1) = 0 \\ p(s_i, d_j, t) \neq 0 \end{cases} \tag{5}$$

The above four cases describe the changes of network behavior on different logical links, case 1 describes the stability of network behavior statistically, and it is the main feature of network behavior. The other three cases describe difference changes of network behavior. Case 2 describes the changes of transmission between network groups, which shows the abrupt change of data transmission between groups. Case 3 and case 4 describe the migration of network behavior among different groups. Those three cases are the key characteristics which should be extracted for abnormal behavior detection.

4.2 Network Abnormal Behavior Detection

The occurrence of network abnormal behavior will lead to changes of network traffic patterns, the distribution characteristics of groups which include hackers and victims will change in a certain degree. When there are massive attacks (such DDOS attacks and worm attacks), the network traffic will have significant changes on the logical link. Furthermore, the behaviors among different groups include victims are will express a certain synergy, this paper users two parameters to measure those large-scale network abnormal behaviors.

We define the similarity between network group behaviors to measure the synergy among network group behaviors. The behavior similarity between group s_i and group s_j is defined as formula (6).

$$r(S_{ij}) = \frac{\sum_{k=1}^{n} w_{ijk}\left(p(s_i, d_k) + p(s_j, d_k)\right)}{\sum_{k=1}^{n} p(s_i, d_k) + \sum_{k=1}^{n} p(s_j, d_k)} \tag{6}$$

The weight coefficient w_{ijk} means $p(s_i, d_k)$ and $p(s_j, d_k)$ are not null at the same time. It to say if there is communication relationship between the source group s_i and

the destination group d_k, then the value of w_{ijk} is 1, otherwise 0. If the similarity between groups is bigger than the threshold, then we can conclude that there is a synergistic phenomenon between network group behavior, and can further analyze the reasons for the phenomenon.

We define the maximum value of network group behavior change as the abnormal degree of network group behavior, which can be used to described the changes of network group behavior on logical links, to describe whether there are large-scale network abnormal behaviors on logical links, such as massive network attacks (DDOS and DOS attacks)

$$c(s_i) = \max_j(|\Delta p(s_i, d_j)|) \tag{7}$$

5 Experimental Results

5.1 Network Traffic Acquisition

The network traffic used in the paper collects from the ingress router of the backbone network of Northwest network center of education network, and the bandwidth is 10 Gbps. Northwest network center of education network is the access point of CERNET for the educational institutions in Northwest provinces, and also the access point for some commercial organizations. There are about 3 million hosts are access to internet through it. The monitored network includes nearly 80 C network segments, the monitoring hosts about 10 thousand, and the web traffic about 8 Gbpm, and the network traffic used in the experiment lasted about 10 h.

5.2 Network Traffic Analysis

According to the definition of network group flow, firstly the network IP addresses should be solved by 24-bit network address prefix, which can form different network user groups, and then form different network links according to the communication between groups, and use the distribution of network traffic on the logical links to describe the characteristics of network group behavior.

We extract a sampling time randomly, the distribution of network traffic on different logical links is shown as Fig. 3, from the figure we can see that the amount of variation of data flow between different groups, and a small number of region flow carries most of network data, there are only account for about 1/10000 of the total traffic in some region flow, which is the phenomenon of heavy tail of network traffic.

Optical flow analysis is widely used in the feature extraction in images [9], the difference between adjacent frames of moving images can reflect the change characteristics of images, which can be used to detect the moving object and extract features. In this paper, we use the distribution which is formed by network traffic in different time to describe the characteristics of network behavior, and the difference between adjacent distribution graphs can describe the dynamic characteristics of network behavior. We take two adjacent frames randomly, and the differences of them are

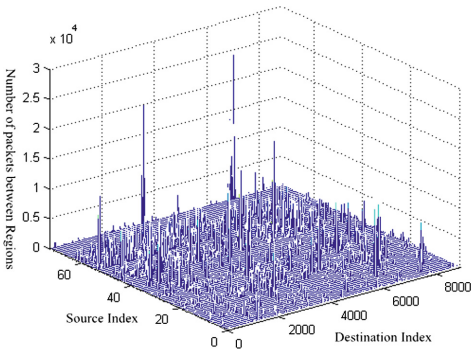

Fig. 3. The distribution of network traffic on the region links

shown as Fig. 4. From Fig. 4, we can see that the network behaviors described by distribution are approximately stable on most logical links of groups, and the dynamic changes are close to zero. But there are a few larger changes of network behavior in a small number of networks, which express the increase and decrease of communication volume among the network groups dramatically. According to the characteristics, we can locate the logical links which have violently changes of network behaviors and network groups, and lay the foundation for the control of network abnormal behavior.

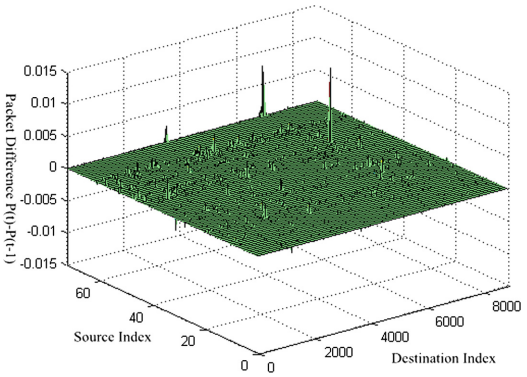

Fig. 4. The difference between adjacent frames of network distribution graph

From the Fig. 4, we can see that there are three peak positions about the difference between adjacent distributions, which descripts the changes of network behavior on three logical links are obviously, according to analysis, we can also find that the changes account for 23.6% of the total data changes, the three logical links are (s_{50}, d_{7145}), (s_{73}, d_{6899}) and (s_{73}, d_{6938}) respectively. The changes of network behavior on the three-logical links lead to the overall changes of network behavior, which is very important information to study network abnormal behaviors. In addition, from Fig. 4, we can also see that the changes in the amount of data are mainly concentrated on the

indexes of three groups, s_{50}, s_{64} and s_{73}, the variability of behavior of the users in the three source groups accounts for 48.2% of the total variability of behavior. The three source groups fluctuate intensely for network behavior, and they are the focus groups which we should monitor.

5.3 Network Abnormal Behavior Detection

According to the change characteristics of the network behavior on the logical link described above, the paper calculates the similarity of network behaviors among network source groups, and the results are shown as Fig. 5.

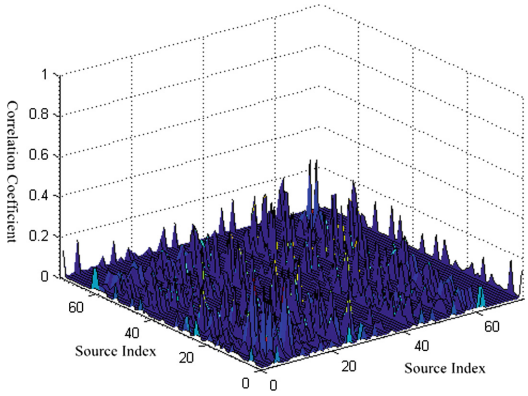

Fig. 5. The similarity of network behavior among source region

From Fig. 5, we can see that the similarity of network behavior among network source group is very low generally. That is to say, the similarity of network behavior among network source group is very small, namely there is no large-scale network collaboration behavior in the network. But if it appears the network collaboration behavior of different groups in network, namely there are the same behavior characteristics in different groups, the hosts in the network groups may be controlled by hackers, who are detecting the network or some other destructive activities for network.

The change degree of the group behavior to measure the changes of network abnormal behavior is shown as Fig. 6. From the figure we can see that the vast majority of network group behavior between adjacent frames is table, but we can also see that there is a group which changes dramatically, denoted by an asterisk. According to the change we can conclude that there are massive network abnormal behavior in the source groups of index numbers 73 and 50. According to analysis, we find that there are DOS attacks on the two logical links at the time. There are a large number of UDP packets on the two logical links flowing to 80 ports, which seriously occupy the network bandwidth.

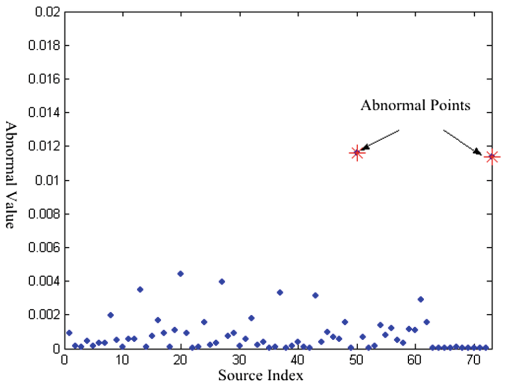

Fig. 6. The change degree of region behavior

6 Summary

As it is very difficult to monitor large-scale network traffic, we proposed the conception of network group behavior flow. From the results of experiment, we can see that the network group behavior flow can describe the characteristics of network traffic effectively, although the method of group flow aggregation can reduce the fine-grained characteristics of some network traffic, the conception of group flow can describe the characteristics of large-scale network flow effectively. The optical flow analysis proposed in the paper can describe the dynamic changes of network group accurately and detect the fluctuation groups of network behavior in real time, and lay the foundation for the control of network abnormal behavior.

Acknowledgments. The research presented in this paper is supported in part by the Natural Science Foundation of China (61502438, 61672026), Natural Science Foundation of Shaanxi Province (2016JM6040), and Chinese Defense Advance Research Program (B0820132036).

References

1. Baldi, M., Baralis, E., Risso, F.: Data mining techniques for effective and scalable traffic analysis. In: IEEE International Symposium on Integrated Network Management, 15–19 May 2005, pp. 105–118 (2005)
2. Zhou, A., Guang, C., Guo, X.: High-speed network traffic measurement methods. J. Softw. **25**(1), 135–153 (2014)
3. Zhang, B., Yang, J., Wu, J.: Survey and analysis on the internet traffic model. J. Softw. **22**(01), 115–131 (2011)
4. Wang, J., Rossell, D., Cassandras, C.G., et al.: Network anomaly detection: a survey and comparative analysis of stochastic and deterministic methods. In: Proceedings of the 52nd IEEE Conference on Decision and Control, Florence, Italy (2013)
5. CISCO NetFlow: Cisco systems, Inc. (2004). http://www.cisco.com/en/US/products/ps6601/products_white_paper09186a00800a3db9.shtml

6. Kim, S.S., Reddy, A.L.N.: A study of analyzing network traffic as images in real-time. In: 24th Annual Joint Conference of the IEEE Computer and Communications Societies, 13–17 March 2005, vol. 3, pp. 2056–2067 (2005)
7. Lakhina, A., Papagiannaki, K., Crovella, M., et al.: Structural analysis of network traffic flows. In: Proceedings of the Joint International Conference on Measurement and Modeling Of Computer Systems, pp. 61–72 (2004)
8. Freedman, M.J., Vutukuru, M., Feamster, N., et al.: Geographic locality of IP prefixes. In: Proceedings of the ACM Internet Measurement Conference Berkeley, CA, pp. 153–158, October 2005
9. Wang, X., Zhang, G.: Research on moving object detection method based on optical flow. J. Comput. Eng. Appl. **40**(1), 43–46 (2004)