Journal of Network and Computer Applications 41 (2014) 15–26

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

# A new connection degree calculation and measurement method for large scale network monitoring

Tao Qin [a,*], Xiaohong Guan [a,b], Wei Li [a], Pinghui Wang [a], Min Zhu [a]

[a] MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] Department of Automation and TNLIST Lab, Tsinghua University, Beijing 100084, China

ARTICLE INFO

*Article history:*
Received 2 May 2012
Received in revised form
8 September 2013
Accepted 7 October 2013
Available online 29 October 2013

*Keywords:*
Abnormal behavior detection
Bi-directional flow
Degree correlation analysis
Renyi entropy
Reversible degree sketch

ABSTRACT

Traffic pattern characteristics monitoring is useful for abnormal behavior detection and network management. In this paper, we develop a framework for connection degree calculation and measurement in high-speed networks. The bi-directional traffic flow model is employed to aggregate traffic packets, which can reduce the number of flow records and capture user's alternation behavior characteristics. The first order connection degree and joint correlation degree are selected as the features to capture the characteristics of traffic profiles. To perform careful traffic inspection and attack detection, not only the abnormal changes of a single traffic feature but also the correlations between the features are analyzed in the new framework. First, the symmetry of in and out connection degrees is analyzed. And we found that incomplete flows are an important information source for abnormal behavior detection. Second, joint correlation degree can characterize the user's communication profiles and their behavior dynamics, which are employed to perform abnormal detection using measurements based on Renyi cross entropy. Finally, the reversible degree sketch is employed for querying abnormal traffic pattern sources for real-time traffic management. The experimental results based on actual traffic traces collected from Northwest Regional Center of CERNET (China Education and Research Network) show the efficiency of the proposed method. The method based on Renyi entropy can detect abnormal changing points correctly. *FNR* of the reversible sketch for locating abnormal sources is below 4% and time complexity is constant and less than 4 s, which is critical for real-time traffic monitoring.



## 1. Introduction

Capturing and analyzing the abnormal traffic is one of the most critical issues in keeping a network under control. This attracts many researchers in recent years. Abnormal behaviors are referred to the behaviors cased by attacks which infect normal operations of the Internet, such as worms and DDoS, etc. Those attacks will cause changes in the flow patterns. Abnormal behavior detection is defined as how to discover those attacks by analyzing the flow patterns using statistical methods. Traffic flow is one of the most important information sources for abnormal behavior detection (Brutlag, 2000). By analyzing the statistical characteristics of the traffic packets (number of bytes, number of packets, etc.) with machine learning and signal processing approaches, many network monitoring methods are successfully developed to discover anomalies due to traffic pattern changes caused by attacks (Roughan et al., 2004; Kim et al., 2004; Barford et al., 2002; Mahoney, 2003; Giorgi and Narduzzi, 2008). Although significant changes of network traffic patterns can be captured by many methods in the existing literatures, it is difficult to detect anomalies which do not cause significant changes in the probabilistic distribution of a single traffic feature, e.g. worm detection at their early propagation age. At the early propagation stage of worms, the infected hosts will send many scanning packets searching for vulnerable targets. Their scanning behaviors will cause changes in the statistical traffic features, including the connection degrees, number of packets, flows, and bytes. On one hand, to control their propagation and to reduce the losses, we must detect the worms at their early propagation stage. However, on the other hand, at the early propagation age, the number of infected hosts is small and abnormal changes of flow features are slight in large-scale networks. If we simply analyze one of those traffic features, it is difficult to detect them. However, if we analyze the related features comprehensively, detection would be easier. Many attacks in the networks today become smarter and tend to gradually change their behaviors to avoid causing sudden changes in traffic patterns. Methods based on obvious changes may lose their efficiency and effectiveness. Furthermore, the rapid increase of bandwidth and number of users with massive flow records cause serious computational difficulties in anomalies detection.

* Corresponding author. Fax: +86 2982664603.
*E-mail address:* tqin@sei.xjtu.edu.cn (T. Qin).

1084-8045/$ - see front matter 
http://dx.doi.org/10.1016/j.jnca.2013.10.008

In this paper, we employ the bi-directional traffic flow model to extract the traffic metrics of large-scale networks and to reduce the number of flow records. It aggregates forward and backward packets generated by the same internal and external hosts during a specified time window $T$ based on the 2-tuples: source and destination addresses. It thus differs from the traditional 5-tuples flow model. The first advantage of this new flow model is that the amount of processed data is greatly reduced compared with individual Netflow models, since the traffic metrics with the same addresses are aggregated. Second, the flow model can reflect the alternation behavior characteristics as the forward and backward traffic packets are aggregated together. The time window $T$ is selected according to the requirement of real-time network monitoring. Based on the operation experiences on CERNET, we select the time window $T$ as 180 s in this paper. The flow model can be described by a directional graph denoted by $G(V, E)$, with $V$ being the set of $N$ nodes and $E$ the set of $M$ edges (Aiello et al., 2005). Our actual traffic collection point is placed at the ingress router of Northwest Region Center of CERNET, which divides the hosts into two parts: inside and outside of the collection point. We could only capture the communication traffic patterns among the users inside and outside of the collection point. The hosts inside of the collection point have internal addresses named as sources while the hosts outside have external addresses named as destinations. The hosts are the set of vertices $V$ and the logic links between the internal and external hosts are the set of edges $E$. We use the IP addresses to identify the hosts without considering the situation of DHCP (Dynamic Host Configure Protocol). Based on this model, two kinds of traffic features are introduced to characterize the traffic patterns. First, the first order connection degree is defined to capture the communication range of the hosts, such as the number of different addresses they access in a time window $T$. The source connection in degree (*SCID*), source connection out degree (*SCOD*), destination connection in degree (*DCID*), and destination connection out degree (*DCOD*) are firstly defined. Those first order connection degrees can capture the users' access regularity. And the definitions per host can greatly reduce the amount of data to be processed and can increase the computational efficiency. Second, we employ the joint correlation degrees ($P_1$, $P_2$, $P_3$, $P_4$) to capture the communication structure. It is the edge-paired production of the network graph, which has been proved very efficient to capture the nature of the network connectivity (Edward, 2006).

We first analyze the symmetry characteristics of the first order connection degree, and then find that most of the in and out connection degrees of specific host are similarly equal to each other. We also find that few users have significantly different number of in and out connection degrees. This is caused by the incomplete connection for the lack of forward or backward packets. The traffic collection point is the unique ingress router of the Northwest Region Center of CERNET, and all the traffic packets should be captured. As the network is used for information exchanges, one normal connection behavior should include both forward and backward packets. Thus, we regard the incomplete flows as an important information source for abnormal detection. To describe users' connection behaviors more specifically, we analyze the joint correlation degree distributions and find that there are four kinds of communication structures. They reflect the users' network access behavior profiles and dynamic changes in the traffic patterns. Based on the above analysis, a systematic measure is employed to perform careful inspection into the traffic patterns and detects attacks by correlating small pattern changes of the related traffic features. There are two correlation categories: (1) the correlation between the related features at a particular time instant; (2) the auto-correlation of the same feature between adjacent time instants. The correlations are quantitatively measured by Renyi cross entropies. If there are no significant changes in the traffic patterns, the Renyi cross entropy approximates zero. If not, the entropy would depart from zero. We can identify the abnormal time points using this method.

We employ the reversible degree sketch to calculate the connection degrees for online application and obtain the detailed IP addresses of abnormal hosts for traffic management. By employing the Chinese remainder theory to design hash functions of the sketch, we only use the information of hash functions to reconstruct the keys without using any keys' information. In other words, it is very efficient because the computational time is constant and is irrelevant to the key space size. Thus, we can infer the keys responsible for obvious changes accurately and efficiently in high-speed networks.

To demonstrate the effectiveness of the new method, several actual traffic data sets collected from the Northwest Regional Centre of CERNET (China Education and Research Network) are analyzed. The abnormal behaviors of data traces are manually found with time consuming and “labor intensive” efforts. The new method is validated with these labeled data traces. The regular entropy, EWMA and relative entropy methods are employed to evaluate its performance. Furthermore, the *joint method* with different sampling rates is selected to evaluate the performance of the reversible sketch. The experimental results show that the new method outperforms other existing methods. Based on correlation analysis of related flow features, Renyi entropy based method detects the abnormal points correctly. Furthermore, *FNR* of the reversible sketch for locating abnormal sources is below 4% and time complexity is constant and less than 4 s, which is critical for real-time traffic monitoring.

The paper is organized as follows. Section 2 briefly describes the related work. Section 3 describes the traffic flow model and the definitions of traffic features. Section 4 presents the method for measuring the characteristics of degree and dynamic changes. Section 5 analyzes the characteristics of the traffic data sets. The results of dynamic change measurement are presented in Section 6. Finally, performance of the reversible degree sketch is analyzed in Section 7. Concluding remarks and future work then follow.

## 2. Related works

Traffic analysis and monitoring are important for Internet management, which are widely studied in recent years. By analyzing the IP address attributes many interesting findings are proposed for abnormal behavior detection. In Kim and Reddy (2005), traffic packets are projected to four matrixes according to different bits of the IP address, and an abnormal detection method for large-scale network is proposed. The structure of addresses contained in IPv4 traffic with different length of prefixes is analyzed and many interesting findings are proposed for traffic measurement and monitoring in Kohler et al. (2006). In addition to analyze the characteristics of IP addresses, many researchers also analyze the statistical characteristics of users' behaviors and perform abnormal behavior detection (Tan et al., 2003; McDaniel et al., 2006; Aiello et al., 2005). The Protocol, Client, Server Port, and total number of packets transferred are analyzed to describe the users' communication patterns as well as to group them into different community of interests (COI). By analyzing the characteristics of the COIs, many abnormal behavior detection methods have been developed (John and Tafvelin, 2007). However, those methods require checking every packet to get the detailed IP addresses, which may affect their efficiency in real time traffic monitoring applications.

To avoid checking every packet and to improve the efficiency, methods analyzing the statistics of the traffic packets (total number of bytes, etc.) are successfully designed to discover anomalies based