Detecting and Identifying Network Anomalies by Component Analysis

LE The Quyen [1], Marat ZHANIKEEV [1], and Yoshiaki TANAKA [1,2]

[1]Global Information and Telecommunication Institute, Waseda University
1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo, 169-0051 Japan
[2]Advanced Research Institute for Science and Engineering, Waseda University
17 Kikuicho, Shinjuku-ku, Tokyo, 162-0044 Japan
E-mail: quyenlt@fuji.waseda.jp, maratishe@asagi.waseda.jp, ytanaka@waseda.jp

Abstract. Many research works address detection and identification of network anomalies using traffic analysis. This paper considers large topologies, such as those of an ISP, with traffic analysis performed on multiple links simultaneously. This is made possible by using a combination of simple online traffic parameters and specific data from headers of selective packets. Even though large networks may have many network links and a lot of traffic, the analysis is simplified with the usage of Principal Component Analysis (PCA) subspace method. The proposed method proves that aggregation of such traffic profiles on large topologies allows identification of a certain set of anomalies with high level of certainty.

Keywords: Network anomalies, Anomaly detection, Anomaly identification, Principal component analysis, Traffic analysis

1 Introduction

Nowadays, computer networks and traffic running through them are increasing at a high pace to meet users' requirement. Beside the major proportion of productive traffic, there are many types of network anomalies. The prominent point of all network anomalies is that they generate abnormal changes in traffic features such as bandwidth, load, and other traffic metrics. In this paper, we concentrate on large network topologies connecting many networks by multiple links. Controlling network anomalies in this scope requires collecting traffic and offline processing data on all network links simultaneously. In our research, we propose to use simple link utilization metrics, i.e. bandwidth (bps), load (packet per second), and counters for selective packets to detect and diagnose network anomalies. This paper applies component analysis and uses subspace method to detect abnormal exhibitions in each link metric. This discovery about anomalous features in network traffic allows us to detect and identify them in a timely manner. The efficiency of this method depends on the data sampling rate which contains the detailed level of network traffic.

2 Traffic and Network Model

Within the scope of this paper, we only address a certain set of network-centric anomalies including: AlphaFlow, Scanning, Neptune, Network outage, and FlashCrowds. Introduction about these anomalies can be found in Wikipedia webpage [1]. Regarding connection establishment behaviour, we noticed that network-centric anomalies fall into 3 categories: point-to-point, point-to-multipoint, and multipoint-to-point. Therefore, in order to decide if an anomaly belongs to any of the 3 categories, we count the number of packets with distinguished source-sockets (address and port) and the number of packets with distinguished destination-sockets appearing on each link. As shown in Table 1, by analyzing link bandwidth, load, the number of distinguished source-sockets and destination sockets to detect abnormal change, we can get significant information to identify network anomalies.

We use OPNET Modeler to simulate a backbone network with 7 router nodes and 8 physical links as shown in Fig. 1. We use most of the common network services to provide normal traffic in this topology including: web browsing, email, database access, file transfer, and voice over IP calls. In order to inspect and to verify the efficiency of the proposed method, we subjectively insert 6 anomalies including UDPFlood, Network outage, FlashCrowds, Portsweep, Neptune and IGMPFlood into the network at various specific times. These anomalies are recreated from packet trace files generated by attacks in real network environment, and then imported and regenerated in OPNET. We run the simulation for 1 day and collect values for each metric from all links at 2 sampling intervals: 30 seconds and 5 minutes.

Table 1. Distribution of values in different features of network anomalies.

Anomaly	BW	Load	S-socket	D-socket	Timespan
AlphaFlow	large	large			short to long
Spoof or distributed flooding, FlashCrowds	large	large	large		long
Network outage	small	small	small		long
Neptune		large	large		long
Scanning		large	large	large	short

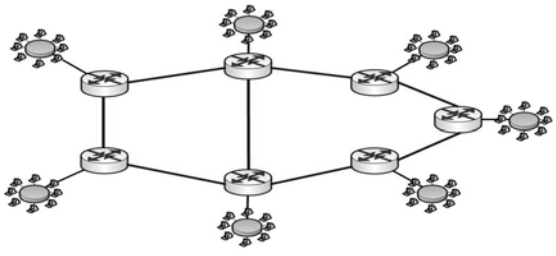


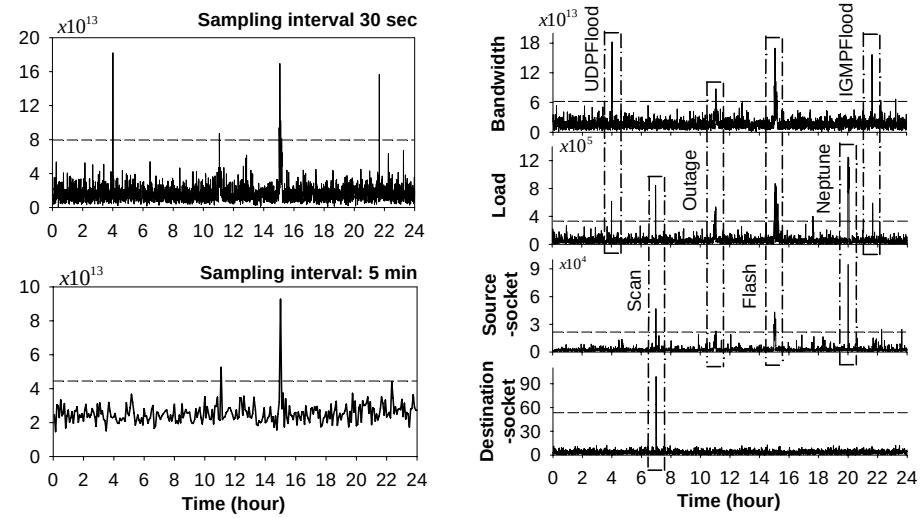
Fig. 1. Network topology.

3 Analysis by PCA Subspace Method

In order to learn the current state of network traffic, we create the state matrix for all links where the rows represent sequential timeline for data collection and the columns represent traffic links. Then, the dimensionality of network traffic state is decided by the number of traffic links, which is usually quite large. Our intention is to detect abnormal values or spikes in metric values of network traffic so we apply PCA subspace method. This method is also used in some other researches for anomaly detection purpose [2], [3]. Detailed explanation can be found in our previous research [4].

PCA is used to explore the intrinsic data dimensionality, and transform the given datapoints into a new space with new coordinates. The dataset will be transformed into a new space generated by PCA where principal components (coordinates) build the traffic normal behaviour and residual components build the anomalous behaviour. Therefore, the PCA space is divided into 2 subspaces: normal (from principal components) and anomalous (from residual components). By calculating the residual vector (constituted by values on all coordinates of anomalous subspace), we can detect anomalies in certain network traffic metric.

Fig. 2(a) shows the detection results by subspace method on the 30-second and 5-minute sampling interval dataset of traffic bandwidth. The thresholds to decide what peaks are anomalies are calculated based on the research on residual components in [5]. According to Fig. 2(a), only the 30-second sampling interval allows trustworthy detection as it successfully detects 4 inserted anomalies. The reason is that PCA is an energy-based analysis, and that it can only detect abnormal variances which are strong enough to create significant change.



(a) Residual bandwidth vector at different sampling intervals

(b) Residual vectors of network link metrics

Fig. 2. PCA subspace method detection results.

The duration of the inserted UDPFlood and IGMPFlood is 1 minute and 30 seconds respectively, so all features in traffic are smoothed out at the end of a 5-minute interval. In contrast, the duration of Network outage and FlashCrowds anomalies are both more than 10 minutes, which ensures that the measurement interval will result in an anomalous reading. This supports the notion that the higher the traffic data sampling rate, the better the detection result.

4 Identifying Anomalies

According to the discussion in Section 2, the anomalous behaviour of distinct types of anomalies has different signatures in each of the four main metrics: bandwidth, load, the number of distinct source sockets and distinct destination sockets. Therefore, we also apply subspace method to the other 3 parameters of all links to detect anomalous patterns in each individual traffic feature. The results are put together for simultaneous analysis as shown in Fig. 2(b). We see that network anomalies exhibit anomalous behaviour differently in each of the 4 traffic metrics. The detection results match the initial comment in Table 1. In case of network outage, even though when it occurs, most of the metrics decrease abnormally but in PCA analysis, such exhibition is still considered as data variance so it will create spikes in residual vectors. This makes the detection result of network outage similar to that of FlashCrowds. Besides, when there is a failure in a part of the topology, network nodes always tend to find substitutive resource to use, so the variance of traffic metrics is not large enough to create major spikes.

5 Conclusion

In this paper, PCA subspace method is applied to detect and to identify network anomalies based on link traffic analysis. With the traffic obtained through simulation experiments, the efficiency of the proposed method in detecting network anomalies is proved. We also address the issue of sensitivity of the method that depends on the rate of traffic sampling. Since PCA is an energy-based analysis, the method is more accurate when anomaly's energy (variability) is comparatively high. It is left for further study to verify the method using traffic from real network environments which allow us to add parametric substantiality to the proposed method.

References

1. Wikipedia webpage link: http://en.wikipedia.org/wiki/Main_Page
2. Wang, W., Battiti, R.: Identifying Intrusions in Computer Networks Based on Principal Component Analysis, Technical Report DIT-05-084, University of Trento (2005)
3. Lakhina, A., Crovella, M., Diot, C.: Diagnosing Network-Wide Traffic Anomalies, Proc. ACM SIGCOMM, Portland, USA (Aug. 2004) 4-6
4. Le, T. Q., Zhanikeev, M., Tanaka, Y.: Component Analysis in Traffic and Detection of Anomalies, IEICE Technical Report on Telecommunication Management, No.TM2005-66 (March 2006) 61-66
5. Jackson, J. E., Mudholkar, G. S.: Control Procedures for Residuals Associated with Principal Component Analysis, Technometrics (1979) 341-349