Detecting Anomalies in Massive Traffic with Sketches

Sirikarn Pukkawanna, Hiroaki Hazeyama, Youki Kadobayashi, and Suguru Yamaguchi
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{sirikarn-p, hiroa-ha, youki-k, suguru}@is.naist.jp

ABSTRACT

Sketches have been considered as an efficient and scalable structure for processing massive data. In this work, we propose a sketch-based method for detecting anomalies in network traffic. The method divides an IP traffic stream into sub-streams using the sketches and detects anomalies in the sub-streams based on a time-frequency analysis of the sub-stream's entropies. The paper shows detection and false positive rates of the method that was evaluated with real-world 150 Mbps traffic collected at the United States and Japan transit link.

Categories and Subject Descriptors

C.2.3 [**Network Operations**]: Network monitoring

General Terms

Algorithms, Security

Keywords

Anomaly detection, sketch, entropy, time-frequency analysis, S-transform

1. INTRODUCTION

As the Internet and traffic data continue to grow exponentially, future traffic analyzers have to support higher data rate. Sketch is an efficient and scalable structure that suites to process massive data stream due to the use of fixed memory size to maintain the processed data. Recently, statistic-based network anomaly detection techniques have been taking the advantage of the sketches for data processing at initial stage before performing anomaly detection stage.

In our previous work [1], we proposed a time-frequency analysis-based method for anomaly detection that utilizes S-transform to convert traffic volume time-series (e.g., packet rate) to time-frequency domain. Frequency changes, referred to as anomalies in this work, in the time-frequency domain are detected by a heuristic-based method.

This paper extends the previous work. More specially, we instead consider traffic entropy time-series due to its capacity to capture more fine-grained traffic patterns than volume-based traffic features. We use sketches to store time-varying entropy values of sub-streams, which are divided from an original traffic stream. S-transform analysis is performed to the sub-stream's entropies to detect anomalies. Analyzing the smaller scale traffic data provides a more elaborate investigation which is likely to increase detection rate.

CFI '14, Jun 18-20 2014, Tokyo, Japan
ACM 978-1-4503-2942-2/14/06.
http://dx.doi.org/10.1145/2619287.2619301

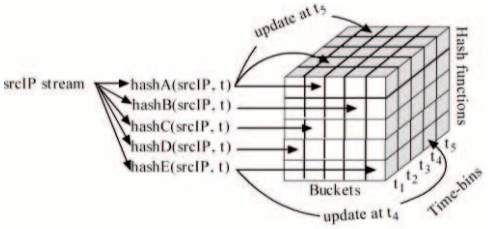

Figure 1. Constructing sketches by five hash functions.

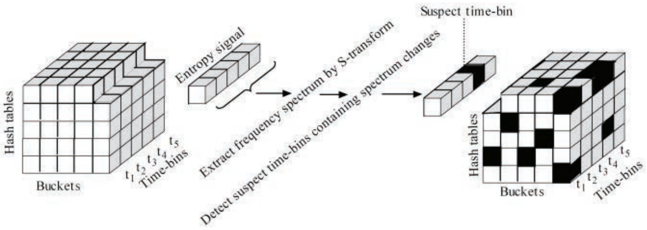

Figure 2. Detecting suspect time-bins from the entropy signal of a sub-stream.

2. PROPOSED ANOMALY DETECTION METHOD

2.1 Sketching Entropies

IP traffic stream is comprised of packets that have four basic attributes: source IP, destination IP, source port, and destination port. In this stage, an attribute stream (e.g., source IP stream) is divided to sub-streams by hashing. More specially, all attributes in a time-bin are hashed individually by different hash functions, and stored in a sketch, which is two-dimensional array. Each row is associated to a hash function and the columns are hash buckets. Each bucket stores the attributes that have hash keys equal to the bucket number. Figure 1 shows the five continuous sketches that are being constructed in five time-bins by five hash functions. The input is the source IP stream and the number of buckets per hash function is five. Next, we compute the Shannon entropies of attributes in each bucket by $H(X) = -\sum_{i=0}^{n} p_i log_2 p_i$, where the p_i is the probability of attribute x_i in the bucket. The p_i is calculated by the frequency of the attribute x_i divided by the frequency of all attributes in the bucket. The reason we consider the entropy instead of volumes of the attributes (e.g., total number of bytes associated with all attributes in a bucket) is that the entropy provides more fine-grain information of traffic data. In the rest of the paper, a vector of the time-varying entropies of a bucket number is called entropy signal.

2.2 Detecting Suspect Time-bins

This stage detects suspect time-bins of each sub-stream. Typically, anomalies are referred to as events that behave differently from major behavior. In others words, anomalies are referred to as changes. In this work, we do not detect changes in the entropy signal of a sub-stream but detect changes in spectrum of the entropy signal using S-transform. The S-

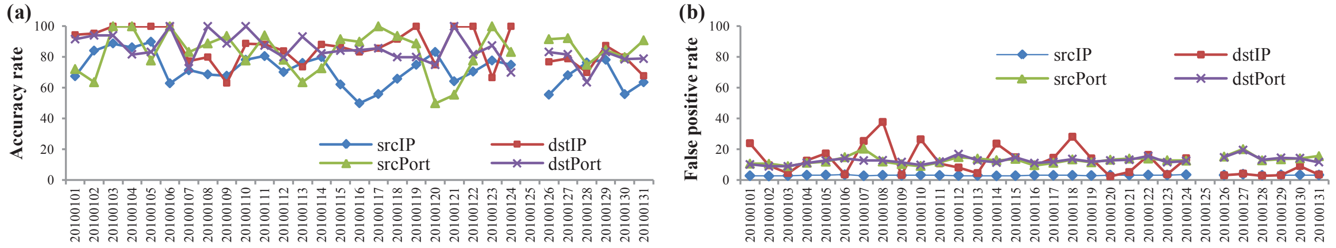

Figure 3. (a) Accuracy rate and (b) False positive rate in detecting anomalous source and destination IPs, source and destination ports in traces of January, 2010

transform is a time-frequency analysis tool like Wavelet transform but it produces an output which is easier for analysis and retains absolute phase information of an input signal.

Firstly, the entropy signal is normalized by subtracting its mean value. Secondly, the S-transform converts the normalized signal and produces a matrix indicating frequency spectrum of the signal including time information. The columns of the matrix represent time-bins corresponding to time-bins of the sketches. The rows represent frequencies and the element is frequency amplitude. In order to determine suspect time-bins, we produce two additional time-series that are obtained by vertically summing all matrix elements in: 1) the upper half, and 2) the lower half of the matrix. Time-bins in the time-series that hold values above a given upper threshold value or below a lower threshold value will be determined as suspect time-bins of the particular sub-stream. Figure 2 depicts the processes of detecting suspect time-bins of a sub-stream. The suspect time-bins are highlighted in black. In order to finish this stage, all entropy signals are examined to find suspect time-bins.

2.3 Finding Intrinsic Culprits

The suspect time-bins contain culprits of anomalies. In order to detect the suspicious culprits, we combine all attributes in the suspect time-bins of each hash function by taking the union. Intrinsic culprits that are hiding among the suspicious culprits are determined by taking the intersection among all suspicious culprits of all hash functions. The attributes in the intersection results are the intrinsic culprits of the traffic stream.

3. PERFORMANCE EVALUATION

As standards of comparison frameworks and labeled traffic datasets for evaluation of anomaly detection methods are lacking, we evaluated the performance of our method with public unlabeled traffic traces from MAWI repository [2] and used detection results of MAWILab [3] as benchmark.

3.1 Traffic Dataset and Anomaly Labels

The MAWI traces [2] organized by WIDE project were used for the evaluation. The trace contains 150 Mbps backbone traffic that is collected daily at *samplepoint-F* of a transit link between the United States and Japan. We evaluated our method with all traces collected in January, 2010 in which each trace contains about 500,000 distinct IPs. The MAWILab [3] is a recent and successive project that is providing anomaly labels for the MAWI traces. They combine the outputs of four anomaly detectors[1] to classify anomalies into three types of records: 1) ANOMALOUS; 2) SUSPICIOUS; and 3) NOTICE. The ANOMALOUS is anomalous traffic with high probability. The SUSPICOUS indicates suspicious traffic that not clearly were identified by the MAWILab classification method. The NOTICE indicates non-anomalous traffic but was reported by at least one of the four detectors. In this work, we compared our results with the ANOMALOUS records.

3.2 Results

For the evaluation, we set the method parameters as follows. The number of hash functions to construct the sketches is three. We used three hash function from [4]. The number of hash buckets and time-bin size were set to 64 and one second respectively. Accuracy and false positive rates of detection were measured. The accuracy rate was computed by the total number of anomalies that were correctly detected by our method divided by the total number of anomalies that were classified by the MAWILab. The false positive rate is the total number of normal instances that were incorrectly detected as anomalies by our method divided by the total number of normal instances in the trace. Figure 3(a) and (b) plot the accuracy and false positive rates in detecting anomalous source and destination IPs, source and destination ports in 30 traces collected in January, 2010. Note that the accuracy and false positive rates of January 25's trace are not shown because the labels for this trace are unavailable. From the plots, we can observe that the overall accuracy rate is above 60%, and in some traces, our method succeeds in detecting anomalies with 100% accuracy. The false positive rate in detecting anomalous source IPs is low and stable at about 3%. The false positive rates in detecting anomalous source and destination ports are about 12%. While the false positive rate in detecting anomalous destination IPs is ambivalent.

4. CONCLUSTION AND FUTURE WORK

In this paper, we proposed a network anomaly detection method based on sketches and S-transform analysis. We evaluated our method with high-speed backbone traces of the MAWI and used the MAWILab labels as benchmark. The overall accuracy rate is above 60% and up to 100% in some traces. The false positive rates are in the range of 3% to 12% except in detecting anomalous destination IPs that is ambivalent. Our future work includes tackling the false positive problem and exploring the effect of parameters (e.g., the number of hash functions).

5. REFERENCES

[1] Pukkawanna, S., Hazeyama, H., Kadobayashi, Y., and Yamaguchi, S. 2013. Building Better Unsupervised Anomaly Detector with S-Transform. In *NSS,* 2013.

[2] MAWI Traffic Archive, http://mawi.wide.ad.jp

[3] MAWILab, www.fukuda-lab.org/mawilab

[4] General Purpose Hash Function Algorithms, http://www.partow.net/programming/hashfunctions

[1] Hough transform, Gamma distribution, Kullback-Leibler divergence, and Principle Component Analysis-based methods