Computer Communications 36 (2013) 575–588

Contents lists available at SciVerse ScienceDirect

Computer Communications

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

ADMIRE: Anomaly detection method using entropy-based PCA with three-step sketches

Yoshiki Kanda [a,*], Romain Fontugne [b], Kensuke Fukuda [b,c], Toshiharu Sugawara [a]

[a] Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, Japan
[b] The Graduate University for Advanced Studies, Tokyo, Japan
[c] National Institute of Informatics/PRESTO JST, Tokyo, Japan

ARTICLE INFO

Article history:
Received 26 February 2012
Received in revised form 20 August 2012
Accepted 6 December 2012
Available online 17 December 2012

Keywords:
PCA
Hash
Sketch
Anomaly detection
Entropy

ABSTRACT

Network anomaly detection using dimensionality reduction has recently been well studied in order to overcome the weakness of signature-based detection. Previous works have proposed a method for detecting particular anomalous IP-flows by using random projection (sketch) and a Principal Component Analysis (PCA). It yields promising high detection capability results without needing a pre-defined anomaly database. However, the detection method cannot be applied to the traffic flows at a single measurement point, and the appropriate parameter settings (e.g., the relationship between the sketch size and the number of IP addresses) have not yet been sufficiently studied. We propose in this paper a PCA-based anomaly detection algorithm called ADMIRE to supplement and expand the previous works. The key idea of ADMIRE is the use of three-step sketches and an adaptive parameter setting to improve the detection performance and ease its use in practice. We evaluate the effectiveness of ADMIRE using the longitudinal traffic traces captured from a transpacific link. The main findings of this paper are as follows: (1) We reveal the correlation between the number of IP addresses in the measured traffic and the appropriate sketch size. We take advantage of this relation to set the sketch size parameter. (2) ADMIRE outperforms traditional PCA-based detector and other detectors based on different theoretical backgrounds. (3) The types of anomalies reported by ADMIRE depend on the traffic features that are selected as input. Moreover, we found that a simple aggregation of several traffic features degrades the detection performance.

1. Introduction

The number of abnormalities in communication network traffic based on both malevolent and benign intentions has been increasing. The former includes network scanning, worm propagation, DDoS, and so forth, which can have detrimental effects on Internet services. The latter includes flash crowds, sudden changes in demand, equipment failures, etc. In order to constantly and safely operate communication networks and to make good use of a limited number of network resources, we need automatic detection methods that can find abnormal events.

Historically, there are two approaches for the automatic detection of anomalous events: misuse detection and anomaly detection. Misuse detection such as snort [2] matches a packet's payload's patterns to those in the predefined database. Even though it can accurately detect anomalous activities, it is unable to detect new types of worms or unknown misuse activities whose payload's patterns are not included in the database. On the other hand, anomaly detection methods using the statistical behavior of the traffic have recently been attracting a lot of researchers' attention since they do not require a predefined database and have the potential to detect new worms under an assumption that those attacks deviate from the normal statistical behavior.

Our focus in this paper is the anomaly detection methods using the statistical behavior of the traffic. We explain several examples of statistical method applied for anomaly detection. An entropy-based approach for anomaly detection [5] computes the entropy of the distribution of packet feature (IP addresses, ports, etc.) and report anomalies if the entropy value deviates from a standard deviation. Entropy based anomaly detection provides more fine-grained insights than the traditional volume based one. ASTUTE [1] defined a model for normal traffic behavior as short-timescale uncorrelated traffic equilibrium. The equilibrium property holds if the traffic flows (a set of packets that share the same values for a given set of traffic features such as source and destination IP addresses, ports, and protocol number) are nearly independent, and is violated by traffic changes caused by correlated flows. ASTUTE detects anomalies based on such equilibrium property assuming that a large number of flows traverses a non-saturated link. A wavelet-based approach [14,15] detect anomalies by utilizing the difference

* Corresponding author. Tel.: +81 8041324020.
E-mail address: y.kanda@isl.cs.waseda.ac.jp (Y. Kanda).

0140-3664/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comcom.2012.12.002

between the time-varying signals of normal traffics and the abnormal network traffics in frequency band on condition that the energy of anomalous traffics is higher than the total energy in certain frequency band. A multi-scale gamma modeling based approach [10,11] approximates traffic using Gamma distribution and traffic that is distant from adaptively computed reference is detected as anomaly. A Kullback–Leibler (KL) approach [19] constructs several kinds of histograms that monitor distinct traffic features by KL divergence to detect prominent change in traffic. A Principal Component Analysis (PCA) based approach [6,4,7–9,21,22] explains the main feature of traffic by dimensionality-reduction and reports the residual traffic as anomaly. PCA is probably the best-known statistical-analysis technique for network anomaly detection. *Defeat* [9] seems to be the most recent and practical PCA based approach because it helps to specify the network-wide anomalies at a per-host granularity by incorporating entropy-based PCA using sketch [13] techniques (random projection to reduce the dimensionality of the data).

Even though we admire the large contribution of *Defeat*, three points still remain to be more closely investigated including the appropriate sketch sizes, which is the IP header information (source/destination IP addresses or ports) we use as the entropy's original traffic, and a capability comparison with other types of anomaly detections using a longitudinal observation. First, *Defeat* insists that large sketch sizes decrease the missed detection rates and increase the additional detection rates. However, no theoretical explanation for this is given and the data sets they use are two backbone's week-long traces for a limited observation period that does not show the growth of the throughput and the number of unique IP addresses on the Internet. We suggest that the number of unique IP addresses as well as the throughput in the trace have a positive correlation with the appropriate sketch sizes. Also, *Defeat*'s impact of the entropy's choice has not yet been examined. They only merged the anomalies detected by the entropy of a 4-tuple (source/destination IP addresses and port numbers). We claim that the entropy of different IP header information captures different types of anomalies, and thus, it should be essential to carry out the study of the types of detected anomalies by using a different choice of entropy. Thirdly, *Defeat* only compared the result with other PCA-based anomaly detectors. To understand the PCAs merit and demerit for anomaly detection, it is necessary for us to compare the detected anomalies of PCA with another type of anomaly detector.

The main contribution of this paper is fourfold. First, we propose ADMIRE, which is a combination of sketches and entropy-based PCA, but is different from *Defeat* in one important respect, it uses three-step sketches to deal with the packet traces measured from a single link. The proposed method using the three-step sketches performs better than the previous two-step sketches in terms of the true and false positive rate. We describe the mechanism and superiority of the three-step sketches in more detail in Section 3.3. Second, we investigate the correlation between the number of unique IP addresses and the appropriate sketch sizes for Internet traffic traces. Consequently, we can observe the positive correlation between them. To the best of our knowledge, this is the first intensive research using a real backbone trace to characterize the correlation between the appropriate sketch sizes for anomaly detection and the number of unique IP addresses. This finding will be helpful for many anomaly detectors using the sketch technique. Third, by evaluating ADMIRE, we revealed that the different entropy time series for PCA anomaly detection captured the different types of anomalies. As consistent with [5], we strongly believe that we should carefully choose the entropy when we use it for anomaly detection. Finally, we compare ADMIRE's detection capability with the gamma [10] and KL [19] methods using nine-year traces. As a result, ADMIRE performs better than the other methods in terms of its detection capability. Since each method detects different types of anomalies, their use in combination would be effective.

2. Related work

Anomaly detection in backbone network traffic has been intensely studied. Out of many different analysis techniques, PCA-based anomaly detection has recently been a hot research topic because of its ability to detect network-wide anomalies by separating the high-dimensional space occupied by a set of network traffic measurements into two distinguishable subspaces corresponding to the normal and anomalous network conditions [7–9,18].

Lakhina et al. first applied PCA to the origin–destination (OD) flows for the structural analysis of network flows [4]. An OD flow consists of all the traffic entering the network from a common ingress point and exiting the network from a common egress point. They show that PCA can decompose the structure of the OD flows into three main constituents: common periodic trends, short-lived bursts, and noise. They have also shown that the OD flows can be accurately modeled in time using a small number (10 or less) of independent components. Ref. [4] had no sooner been published than the authors also applied PCA to the anomaly detection of OD-flows [7]. The information in Ref. [7] stated that they could detect and identify the anomalous OD-flows that span multiple network links using the time series of the packet count and size as the input into PCA. Lakhina et al. [6] suggests that the entropy time series of traffic features such as IP addresses and ports are better than the packet count and size for the accurate anomaly detection. Li et al. incorporated these works with sketch in order to detect and identify anomalous IP-flows that are more fine-grained than the OD-flow using the entropy time series of traffic features [9]. Liu et al. [18], on the other hand, stated it improved the time complexity of PCA by using the variance estimation achieving a logarithmic running time and space over the traffic streams in the sliding window model using theoretical guarantees. However, their works did not evaluate an important parameter, the number of normal components called htop_kg, which we explain in the Section 3.1 even though Ringberg et al. suggest that PCA-based anomaly detection should be sensitive in top_k [8].

We proposed a packet count-based PCA anomaly detector and approached the top_k's sensitivity problem by using a cumulative proportion-based decision in our previous work [16]. In Ref. [16], we insisted that the adaptive decision of top_k based on the cumulative proportion of the principal components outperforms the fixed decision of top_k proposed in [9]. Even though Refs. [7–9,18] takes advantage of PCAs ability to find anomalies that span multiple links, we conversely enabled PCA to work on single link packet-based traces. Thus, we can evaluate PCA for anomaly detection using much more accessible single link traces than multiple link traces.

In ADMIRE, the input to PCA is not a packet count time series as used in Ref. [16], but an entropy time series, which enables us to evaluate the impact of the entropy-metric's choice for longitudinal observation. We can give insight into the question of what is the most appropriate choice for PCAs input for anomaly detection. Furthermore, we could evaluate the correlation of the appropriate sketch size and the number of IP addresses by making good use of the longitudinal observation of Internet traffic.

3. Methodology

3.1. Principal component analysis (PCA) and subspace method

Principal Component Analysis (PCA) is a famous coordinate transformation technique to explain the characteristics of high