Applied Soft Computing 34 (2015) 513–525

Contents lists available at ScienceDirect

Applied Soft Computing

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

Autonomous profile-based anomaly detection system using principal component analysis and flow analysis

Gilberto Fernandes Jr.[a,*], Joel J.P.C. Rodrigues[a,b], Mario Lemes Proença Jr.[c]

[a] *Instituto de Telecomunicações, University of Beira Interior (UBI), Covilhã, Portugal*
[b] *University of Fortaleza (UNIFOR), Fortaleza, Brazil*
[c] *Computer Science Department, State University of Londrina (UEL), Londrina, Brazil*

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 18 October 2014
Accepted 12 May 2015
Available online 21 May 2015

Keywords:
Network management
Traffic characterization
Anomaly detection
Traffic characterization
Principal component analysis
Flows

ABSTRACT

Different techniques and methods have been widely used in the subject of automatic anomaly detection in computer networks. Attacks, problems and internal failures when not detected early may badly harm an entire Network system. Thus, an autonomous anomaly detection system based on the statistical method principal component analysis (PCA) is proposed. This approach creates a network profile called Digital Signature of Network Segment using Flow Analysis (DSNSF) that denotes the predicted normal behavior of a network traffic activity through historical data analysis. That digital signature is used as a threshold for volume anomaly detection to detect disparities in the normal traffic trend. The proposed system uses seven traffic flow attributes: bits, packets and number of flows to detect problems, and source and destination IP addresses and Ports, to provides the network administrator necessary information to solve them. Via evaluation techniques performed in this paper using real network traffic data, results showed good traffic prediction by the DSNSF and encouraging false alarm generation and detection accuracy on the detection schema using thresholds.

1. Introduction

Nowadays, all sorts of networks are daily targets of attacks and malicious activities that seek to interrupt or disable Internet traffic and services, threatening their availability and operability [1,2]. For instance, Distributed Denial of Service (DDoS) attacks can lead to a serious server overload, by congesting a network with unwanted traffic and requests. Worms, spam distribution, spoofing and cybercrime are other examples of threats that might harm computer networks. However, it is not only attacks that affect the normal network operation. Since networks are growing in size and complexity, problems such as server crash, bugs, link congestion, software failures, and traffic randomness may generate noise on the statistical patterns of the network flow [3–6].

Traffic anomalies are unexpected events in traffic flows that deviate from what is considered as normal. Volume anomalies refer to spikes in the time series of traffic data, usually caused by flash crowds and outage events, which can lead to traffic congestions, reducing the network throughput and increasing network delay. Rapid and accurate anomaly detection and tracing are critical to the efficient operation of large computer networks [7,8].

Therefore, it is necessary to constantly monitor the traffic behavior of a network for early detection of these disparities in traffic flows, so that measures can be taken quickly to solve the problem. However, complete network monitoring and analysis is a tough task to be performed manually by a network administrator in network systems that are growing in size and complexity. The high speeds and large number of links and segments to track make the task even more difficult. Therefore, a monitoring system for anomaly detection must be autonomous and perform proactive network management.

Anomaly detection can be classified in two ways: signature-based, in which prior knowledge about the characteristics of each kind of anomaly is used; and profile-based, which presents a history of the normal network behavior through a network profile and treats any activity deviating from it as a possible intrusion [9]. Although signature-based methods have been widely investigated in the literature, they have a clear drawback. It is prerequisite that anomaly signatures are known in advance, hampering the recognition of new anomalies. Also, signature-based methods can be avoided by malicious sources by tampering anomaly signatures. In contrast, a profile-based system creates a baseline profile of the

* Corresponding author at: Rua Marquês d'Ávila e Bolama, 6201-001, Portugal. Tel.: +351 910438982.

E-mail addresses: gil.fernandes6@gmail.com (G. Fernandes Jr.), joeljr@ieee.org (J.J.P.C. Rodrigues), proenca@uel.br (M.L. Proença Jr.).

http://dx.doi.org/10.1016/j.asoc.2015.05.019
1568-4946/

normal network activity, eliminating the need of prior knowledge about the nature and properties of anomalies. This trait leads to some advantages: the possibility of discovering new and unforeseen types of anomalies; the detection of insider attacks; and also makes it difficult for an attacker to know with conviction what malicious action it can carry out without being detected by the system [9,10]. Thus, this researches' proposal is to create an autonomous profile-based monitoring system capable of identifying the normal network behavior by adopting an efficient method for traffic characterization in order to create a baseline profile of normal traffic to discover possible anomalies in the traffic.

The proposed system presented in this paper is called PCADS-AD (principal component analysis for digital signature and anomaly detection), and it is divided into two steps: traffic characterization and anomaly detection.

A profile-based detection system has its efficiency fully related on traffic behavior characterization. Proença et al. [38] and Assis et al. [33] observed that the traffic network is currently composed of cycles consisting of bursts which have particular characteristics directly affected by workdays and user access period. Therefore, the recognition of these behaviors and their individualities is the core to create the network profile called Digital Signature of Network Segment using Flow analysis (DSNSF). Such signature is responsible for harboring information about the normal traffic behavior, being adopted for anomaly detection by recognizing behavioral deviances which clashes from the usual. In PCADS-AD, traffic characterization is performed by using principal component analysis (PCA) as a mechanism analyze historical input data from network activity, identify the most relevant traffic time intervals amongst the data set, and then reduce them so that this new set can efficiently represent the regular behavior of a network segment.

In PCADS-AD anomaly detection phase, abnormal events are detected based on the DSNSF, which acts as a threshold to generate alarms. Aiming to minimize false alarm generation, information extracted from the principal component analysis performed during the traffic characterization phase is used. This produces confidence bands for the DSNSF, restricting an interval where deviations are considered normal. In addition, the system is able to provide qualitative information about the anomalous time interval, helping network administrator in finding the source of problems, the targets, and its magnitude, in order to solve/control the issue.

The entirety of this research is accomplished through the analysis and extraction of seven IP flow features present in a historical database from State University of Londrina network traffic: the quantitative attributes (bits, packets and number of flows transmitted per second); and the qualitative attributes presented in the IP packet header (source and destination IP addresses and source and destination TCP/UDP Ports). Flow patterns can be extracted by analyzing these statistics, so that changes can be identified in the normal flow patterns caused by abnormal behaviors. Also, it is well-known that multi-dimensional monitoring enables more effective analysis than using only one attribute. According to Zhou et al. [11], single dimension monitoring of a network has become increasingly less effective due some drawback, such as the use of stealth attacks or the inability to identify interesting patterns due to a lack of detail. A multi-dimensional analysis is import since certain kinds of anomalies cause variation on more than one flow attribute. For instance, according to [12,13], the affected attributes for DoS/DDoS are packets and number of flows, while a Flash Crowd anomaly affects bits, packets and number of flows.

The main contributions of this paper consist in generating a digital signature by using principal component analysis in an unusual way than the PCA from the literature, in order to describe the normal behavior of a network segment, and then using it as the basis for anomaly detection. To evaluate the proposed system, a variety of tests were performed using real data from a large university network.

The paper is organized as follows. In Section 2, a related research under the subject of anomaly detection is presented. Section 3, the proposed anomaly detection system PCADS-AD is detailed. In Section 4, the results using evaluation metrics and real traffic data are presented and discussed. And Section 5 provides conclusions and final considerations.

2. Related work

Security in computer networks is an important and vast research topic in the Network Management area. Today, there are many different kinds of anomaly and intrusion detection methods that use all sorts of algorithms and techniques. Xu [14] introduces a novel sequential anomaly detection method based on temporal-difference (TD) learning, and also using reward functions designed in Markovian modeling of sequential data. Lin et al. [15] make use of Support Vector Machine (SVM), Decision Tree (DT), and Simulated Annealing (SA) in order to propose an intelligent anomaly intrusion detection algorithm with feature selection and decision rules. In [16], the authors aim to improve anomaly detection effectiveness by using a new flow-based sampling technique, focusing on the selection of small flows which, the author claims, are usually the source of malicious traffic. Results showed that the detection rate is significantly improved when a sampled technique is used instead of an un-sampled one. In [17] the authors perform a four-class signal analysis of network traffic anomalies using IP Flow and SNMP measurements and wavelet filters to expose both the normal and anomalous traffic. In [18], a novel framework for intrusion detection is proposed based on data mining techniques and fuzzy association rules for building classifiers. And finally, a new version of particle swarm optimization (PSO) algorithm, called the simplified swarm optimization (SSO), is applied to create a hybrid network intrusion detection system with a local search scheme for mining intrusion behaviors [19].

Principal component analysis (PCA) is a widely used technique for anomaly detection in computer networks. Pioneer in this subject, Lakhina et al. [20] addresses the anomaly diagnosis problem in network wide-traffic by using PCA to efficiently separate traffic measurements into normal and anomalous subspaces. The anomaly detection method proposed by Pascoal et al. [21] uses a robust PCA detector combined with a robust feature selection algorithm to obtain adaptability to different network environments and conditions. Also, this robust PCA approach does not require having a perfect ground-truth for training, which is one of the limitations of standard PCA discussed in [22]. In [23], the authors propose ADMIRE, which is a combination of three-step sketches and entropy-based PCA, that results in better true and false positive rates, while it is possible to capture different types of anomalies due to the different entropy time series for PCA.

Regarding profile-based anomaly detection methods, Jiang et al. [24] aim to improve the network anomaly detection capabilities by using a methodology for traffic prediction based on frequency domain traffic analysis and filtering. Basically, the traffic is separated into the baseline component (low frequency and non-stationary traffic) and the short term component (most dynamic part), where the traffic of each part is predicted separately using the ERAN algorithm and ARMA model, respectively. Thus, the total predicted traffic is obtained by the combination of both predicted individual parts. Another approach using traffic prediction is proposed in [10], called sketch-based change detection. It detects traffic anomalies by developing a model of normal behavior based on past traffic history and looking for significant changes in short-term behavior that are inconsistent with it. Furthermore, Amaral