Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 38–47

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Mining Host Behavior Patterns From Massive Network and Security Logs

Jing Ya[1,2,3], Tingwen Liu[1,2,3], Quangang Li[1,2,3], Jinqiao Shi[1,2,3], Haoliang Zhang[1,2,3], Pin Lv[1,2,3], and Li Guo[1,2,3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] National Engineering Laboratory for Information Security Technologies, Beijing, China
{yajing, liutingwen, liquangang, shijinqiao, zhanghaoliang, lvpin, guoli}@iie.ac.cn

Abstract

Mining host behavior patterns from massive logs plays an important and crucial role in anomalies diagnosing and management for large-scale networks. Almost all prior work gives a macroscopic link analysis of network events, but fails to microscopically analyze the evolution of behavior patterns for each host in networks. In this paper, we propose a novel approach, namely Log Mining for Behavior Pattern (LogM4BP), to address the limitations of prior work. LogM4BP builds a statistical model that captures each host's network behavior patterns with the nonnegative matrix factorization algorithm, and finally improve the interpretation and comparability of behavior patterns, and reduce the complexity of analysis. The work is evaluated on a public data set captured from a big marketing company. Experimental results show that it can describe network behavior patterns clearly and accurately, and the significant evolution of behavior patterns can be mapped to anomaly events in real world intuitively.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Behavior Pattern, Network Management, Log Mining

1 Introduction

Network and security logs, such as network health data and network flow data, carry important information for diagnosing network anomalies (*e.g.* port scans, denial of service attacks). Network managers usually achieve troubleshooting depending on analyzing massive logs. Enterprise networks usually contain various network devices and security services (*e.g.* firewalls, routers and network monitoring programs) from different vendors. Thus, large amounts of logs that record hosts' behaviors are generated every day, which can be used to analyze the root cause of network anomalies.

However, analyzing network and security logs becomes an extremely maddening task, because of the following reasons. First, massive logs are generated every day, due to the increasing number of network devices and security services deployed and the increasing network bandwidth.

1877-0509
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.072

Second, real-world logs are often multivariate and evolve over time. Third, logs that record the same hosts' behaviors may span across multiple network layers, services and protocols. Last but not the least, network anomalies usually consist of several subtle behaviors scattered in large amounts of normal logs.

A key step of analyzing such massive and complex logs is to mine the correlation of network behaviors, namely behavior patterns in this paper. Much work have been proposed to address this maddening step [13] [14]. They give a macroscopic analysis to describe network event patterns, for example, using tensor (*sourceIP, destinationIP, Timestamp*). However, they fail to make a microscopic analysis to get the evolution of behavior patterns for each host in networks. We argue that making a microscopic analysis for each host is very important and crucial in understanding the root cause and impacts of network anomalies, because it provides the evolution trend and comparability of behavior patterns and can quickly locate anomalies.

To address this problem, we present an approach named Log Mining for Behavior Patterns (LogM4BP) that can accurately learn and identify the evolution of network behavior patterns for each host from massive and heterogeneous network and security logs. The main idea is to mine every host's behavior patterns in each time window with the help of behaviors of all hosts in a long period of time. We first calculate the feature-pattern matrix W from the massive logs of all hosts, where in W each pattern is a set of features (*e.g.* CPU utilization) extracted from logs. Then we factorize feature-host matrix A^k in the k-th time window into the multiplication of the feature-pattern matrix W and a pattern-host matrix H^k. Matrix H^k represents the patterns of each host in the k-th time window. After calculating all pattern-host matrixes of each time window, we can observe the evolution of each host's network behavior patterns through the measure of dispersion by using the expected value and standard deviation of $\{H^k|k \in [1,t]\}$.

We evaluate the LogM4BP approach on a public data set labelled with the ground truth. The data set is captured from a big marketing company and provided by VAST 2013 Mini Challenge 3. Experimental results show that our work can describe network behavior patterns clearly and accurately, and detect all the labelled and confirmed anomaly events.

The rest of the paper is organized as follows. Section 2 summarizes the related work. We give a detailed description of our behavior mining approach LogM4BP in Section 3. Implementation details and experimental results for this work are shown in Section 4. Finally, Section 5 concludes the paper's work.

2 Related Work

There are many commercial and open-source network management systems (NMSs) for recent complicated network security analytics and management [1, 2, 3]. They visualize various metrics of network security elements, and raise alarms based on predefined rules. However, these rules often indicate obvious and severe network statuses based on domain knowledge, and cannot describe subtle behavior patterns.

A lot of research work has been proposed in the areas of network anomalies diagnosis and localization by analyzing various network and security logs, especially using machine learning.

Yamanishi *et al.* [22] proposed a technique to detect system failure from server syslog using a mixture of Hidden Markov Models. Bahl *et al.* [8] developed an system named Sherlock, which learns a dependency graph between multilevel resources in enterprise networks. G-RCA [12] identified the root cause of the problem by matching a current event to a predefined decision tree, which represents the causal relationships of the network events. Orion [21] extracted the causal relationship among application traffic by analyzing each traffic delay. Meta [19]