available at www.sciencedirect.com

ScienceDirect

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

Computers & Security

An incremental frequent structure mining framework for real-time alert correlation

Reza Sadoddin[*], *Ali A. Ghorbani*

Network Security Laboratory, University of New Brunswick, 550 Windsor St., Fredericton, New Brunswick, Canada E3B 5A3

ARTICLE INFO

Article history:
Received 19 December 2007
Received in revised form
7 October 2008
Accepted 26 November 2008

Keywords:
Online alert correlation
Intrusion detection
Network security
Stream data mining
Graph mining
Attack scenario extraction
False positive analysis
FP tree
DDoS attack
Tilted-time window model

ABSTRACT

With the large volume of alerts produced by low-level detectors, management of intrusion alerts is becoming more challenging. Manual analysis of a large number of raw alerts is both time consuming and labor intensive. Alert Correlation addresses this issue by finding similarity and causality relationships between raw alerts to provide a condensed, yet more meaningful view of the network from the intrusion standpoint. While some efforts have been made in the literature by researchers to find the relationships between alerts automatically, not much attention has been given to the issue of real-time correlation of alerts. Previous learning-based approaches either fail to cope with a large number of generated alerts in a large-scale network or do not address the problem of concept drift directly.

In this paper, we propose a framework for real-time alert correlation which incorporates novel techniques for aggregating alerts into structured patterns and incremental mining of frequent structured patterns. Our approach to aggregation provides a reduced view of developed patterns of alerts. At the core of the proposed framework is a new algorithm (FSP_Growth) for mining frequent patterns of alerts considering their structures. In the proposed framework, time-sensitive statistical relationships between alerts are maintained in an efficient data structure and are updated incrementally to reflect the latest trends of patterns.

The results of experiments conducted with the DARPA 2000 dataset as well as artificial data clearly demonstrate the efficiency of proposed techniques. A promising reduction ratio of 96% is achieved on the DARPA 2000 dataset. The running time of the FSP_Growth algorithm scales linearly with the size of artificial datasets. Moreover, testing the proposed framework with alert logs of a real-world network shows its ability to extract interesting patterns among the alerts. The ability to answer useful time-sensitive queries regarding pattern co-occurrences is another advantage of the proposed method compared to other approaches.

1. Introduction

The significant increase of our everyday life dependency on the Internet-based services has intensified the importance of survivability of networks. Intrusion detection is one of the major techniques for protecting information systems. According to the results of the 2006 CSI/FBI Crime and Computer Security Survey, an intrusion detection system is the fifth most widely used security technology among the respondents (69% of surveyed organizations use IDSs as a security measure).

Intrusion detection has been an active research area for over 20 years since it was first introduced in the 1980s. Intrusion detection systems can be roughly classified as

[*] Corresponding author.
E-mail addresses: resa.sadoddin@unb.ca (R. Sadoddin), ghorbani@unb.ca (A.A. Ghorbani).

0167-4048/$ – see front matter Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2008.11.010

anomaly-based detection and signature-based detection systems. When an intrusion detection system learns the normal behavior of the system or the network it monitors, it is categorized as an anomaly-based IDS. An anomaly is reported when the monitored behavior deviates significantly from the normal profile. A signature-based (misuse) detection approach, on the other hand, uses information about the known attacks and detects intrusions based on matches with existing signatures. These well-known intrusion detection techniques have their own strengths and weaknesses. For example, signature-based intrusion detection has a lower false positive rate but it is intended for detecting known attacks. Anomaly-based detection has the potential to detect novel attacks, but at the same time it suffers from a high false positive rate. Moreover, it is very hard to define normal behavior for a system.

In order to compensate for limitations of individual intrusion detection systems, sensors with various detection mechanisms are deployed in different locations of the protected network. In addition, other security tools such as firewalls, antivirus utilities and file integrity checkers are employed in order to provide a more detailed view of the security status of the network. Managing raw alerts generated by various sensors are becoming more significant to intrusion detection systems as more sensors with different capabilities are distributed spatially in the network. Higher level management is required for analyzing low-level alerts produced by these devices before reporting them to the next layer. This is needed for several reasons. First, it is not easy to locate the source or target of the intrusion or fault in the network by looking into low-level alerts. Secondly, the low-level sensors consider raw alerts in isolation and raise alarms for each of them, without considering logical connections between alerts. Also, any automatic response would be inefficient with these raw alerts as input. In addition, in a typical environment there are a lot of false positive alerts reported by traditional IDSs, which are mixed with true positive alerts.

Alert Correlation provides the system with automatic analysis of alerts. An alert correlation module not only saves a lot of time on the administrative side, but also helps to deal with potential threats against the network more efficiently. Previous works in this area have addressed different aspects of alert correlation including filtering false positives, aggregating similar alerts into clusters, correlating alerts based on their relationship in an attack scenario, and prioritizing alerts based on their severity.

Two main approaches have been used in the literature for correlating alerts in attack scenarios. In the first category of works, the relationships between alerts are hard-coded in the system. These methods are limited to the knowledge base of the system and cannot correlate alerts of unseen attacks. To overcome this problem, machine learning and data mining techniques have been used to extract relationships between alerts after an initial period of training. In this category of works, co-occurrence of alerts within specified time windows is used as an important feature for the statistical analysis of alerts. This involves pair-wise comparison between alerts since every two alerts are candidates to be correlated in a newly emerging pattern. Pair-wise comparison between alerts poses serious challenges on learning-based correlation techniques in large-scale networks, in which a throughput of 5000 alerts per minute is expectable.

In this paper, we address the problem of real-time correlation of alerts in a learning-based approach. The contributions of our work can be summarized as follows:

- We propose a framework for real-time correlation of alerts based on their frequency of co-occurrences. The proposed framework provides a means of processing a stream of alerts continuously, maintaining their correlation significance with respect to the latest changes, and answering some interesting queries about patterns.
- A method for dynamic creation and generalization of alerts into clusters (patterns) is proposed. Abstract signatures of patterns allow us to detect some similar behavior of alerts which otherwise would have been detected as different patterns.
- We propose a novel Frequent Structure Mining technique for extracting patterns considering their structures. The proposed method not only provides more accurate frequency analysis of patterns, but also gives the exact structure of the extracted patterns within the network.

The rest of this paper is organized as follows: In Section 2, some of the related works in alert correlation are reviewed. Section 3 provides the details of the proposed framework for incremental mining of frequent alert patterns. In Section 4, we present results of the experiments that were conducted for evaluating the proposed techniques and framework with the DARPA 2000 dataset, a synthetic graph data generator and a real-world dataset. Finally, the conclusions and some suggestions for future work are given in Section 5.

2. Related work

In the following, we review the related work in the literature, which address aggregating alerts into clusters, correlating alerts into attack scenarios, and filtering false positive alerts.

The purpose of *aggregation* is to group all similar alerts together. During aggregation, alerts are put into a group based on the similarity of their corresponding features. The most common attributes of alerts are *Source IP, Destination IP, Source Port, Destination Port, Attack Class* and *Timestamp*.

In the technique proposed by Valdes and Skinner, 2001, alerts are grouped with each other based on their overall similarity. This overall similarity of two alerts is defined based on their similarities on the corresponding features. The proposed technique, even though it provides a basic probabilistic model for measuring similarities between alerts, has the drawback of relying on expert knowledge for specifying the similarity degree between attack classes in a two-dimensional matrix.

A clustering technique is proposed by Julisch, 2003 for grouping all the alerts which share the same *root causes*. Central to the clustering technique proposed by Julisch are the hierarchy structures (*generalization hierarchies* as they are called), which decompose the attributes of the alerts from the most general values to the most specific ones. The generalization hierarchies are later used for measuring the distance