Chapter 34
A Hybrid Technique Using PCA and Wavelets in Network Traffic Anomaly Detection

Stevan Novakov
Carleton University, Canada

Ioannis Lambadaris
Carleton University, Canada

Chung-Horng Lung
Carleton University, Canada

Nabil Seddigh
Solana Networks, Canada

ABSTRACT

Research into network anomaly detection has become crucial as a result of a significant increase in the number of computer attacks. Many approaches in network anomaly detection have been reported in the literature, but data or solutions typically are not freely available. Recently, a labeled network traffic flow dataset, Kyoto2006+, has been created and is publicly available. Most existing approaches using Kyoto2006+ for network anomaly detection apply various clustering techniques. This paper leverages existing well known statistical analysis and spectral analysis techniques for network anomaly detection. The first popular approach is a statistical analysis technique called Principal Component Analysis (PCA). PCA describes data in a new dimension to unlock otherwise hidden characteristics. The other well known spectral analysis technique is Haar Wavelet filtering analysis. It measures the amount and magnitude of abrupt changes in data. Both approaches have strengths and limitations. In response, this paper proposes a Hybrid PCA–Haar Wavelet Analysis. The hybrid approach first applies PCA to describe the data and then Haar Wavelet filtering for analysis. Based on prototyping and measurement, an investigation of the Hybrid PCA–Haar Wavelet Analysis technique is performed using the Kyoto2006+ dataset. The authors consider a number of parameters and present experimental results to demonstrate the effectiveness of the hybrid approach as compared to the two algorithms individually.

INTRODUCTION

The way networks are being used is rapidly changing and a by-product of this change is that the types of computer attacks are rapidly evolving. For example, malicious attacks are no longer limited to desktop computer viruses, but can target a network itself (Estevez-Tapiador et al., 2004). These attacks are designed to create failures in the system. Depending on the network, these failures

DOI: 10.4018/978-1-4666-7456-1.ch034

can cause mild inconvenience, loss of productivity, loss of economic activity, or even, loss of public well being. This paper applies statistical analysis and spectral analysis techniques to network traffic data in order to identify potential malicious network attacks. Specifically, this paper focuses on a *hybrid technique* to provide information to a network operator such that the source of malicious behavior can be isolated.

There is a need for effective and scalable approaches to maintain network stability and to detect anomalous network traffic behavior created by attacks. This need is increasingly being addressed through the use of flow-based protocols such as Cisco's NetFlow protocol (Cisco, 2012). This protocol resides on routers and each packet that passes through is examined for a set of IP packet attributes. The output of NetFlow is a multi-tuple record, called *a flow*. Some core features of a flow are: Source IP address, Destination IP address, total bytes, etc. NetFlow does not indicate whether a flow is a part of abnormal or malicious behavior (NetFlow, 2012).

Current anomaly detection approaches can be classified into two main categories: knowledge base approaches to identify attacks through patterns for signatures (Bro Secucity, 2012) and approaches to detect patterns which do not conform to expected behavior (Campos & Milenova, 2005). Inspecting individual signatures or traces of known hazards based on a knowledge base is time consuming and inefficient. Furthermore, the turnaround from discovery to updating the knowledge base can be extensive. The second type for anomaly detection is not dependent on an existing knowledge base and identifies potential network threats by finding *deviations* from normal behavior. Some statistical models and signal processing algorithms have been used for this purpose. These methods can be applied relatively quickly to create relationships and discover patterns from a range of data types and sizes. However, a comprehensive anomaly detection system will require a significant amount of human expertise (Campos & Milenova, 2005).

This paper proposes a *hybrid* solution for network anomaly detection based on statistical and spectral analysis techniques, which provides the network administrator time slices containing potential network traffic anomalies. To the best of our knowledge, no such hybrid techniques have deployed by systems described in literature.

The statistical analysis studied is a modified or time shifted Principal Component Analysis (PCA) technique to determine abnormal network behavior (Brauckhoff, Salamatian, & May, 2009). Components are extracted by comparing feature data similarity. A ranked subset of components, selected by comparing the sparsity of projected data, is used to create a subspace that describes anomalous behavior. Time grouped data is projected onto this space and spectral analysis is applied. The feature with the most spread out data is considered in spectral analysis portion.

The spectral analysis technique adopted is Haar Wavelet decomposition (Barford, Kline, Plonka, & Ron, 2002). This type of wavelet decomposition uses a Haar basis function to decompose the input dataset set into core time functions. *Thresholds* are applied to remove noise and highlight network traffic anomaly characteristics. The signal is reconstructed and a weighted score to describe the magnitude of fluctuations within each time slice is calculated. A high score represents a large change in a time window and suggests to the network administrator that abnormal and potentially malicious behavior is present.

This paper proposes a hybrid technique and illustrates that it is more accurate and informative compared with the statistical and spectral analysis techniques independently. This hybrid approach examines network traffic data grouped and *summarized* into time bins. Next, it applies PCA based statistical analysis to reduce the complex nature of the network traffic. Subsequently it employs spectral analysis to determine time slices of interest where there is a change in network traffic