available at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Network anomaly detection through nonlinear analysis

*Francesco Palmieri**, *Ugo Fiore*

Università degli Studi di Napoli Federico II, CSI, Complesso Universitario Monte S. Angelo, Via Cinthia, 80126, Napoli, Italy

ARTICLE INFO

Article history:
Received 10 November 2009
Received in revised form
9 April 2010
Accepted 9 May 2010

Keywords:
Anomaly detection
Recurrence quantification analysis
Nonlinear analysis
Non-stationarity
Support-vector machines

ABSTRACT

Nowadays every network is susceptible on a daily basis to a significant number of different threats and attacks both from the inside and outside world. Some attacks only exploit system vulnerabilities and their traffic pattern is undistinguishable from normal behavior, but in many cases the attack mechanisms combine protocol or OS tampering activity with a specific traffic pattern having its own particular characteristics. Since these traffic anomalies are now conceived as a structural part of the overall network traffic, it is more and more important to automatically detect, classify and identify them in order to react promptly and adequately. In this work we present a novel approach to network-based anomaly detection based on the analysis of non-stationary properties and "hidden" recurrence patterns occurring in the aggregated IP traffic flows. In the observation of the above transition patterns for detecting anomalous behaviors, we adopted recurrence quantification analysis, a nonlinear technique widely used in many science fields to explore the hidden dynamics and time correlations of statistical time series. Our model demonstrated to be effective for providing a deterministic interpretation of recurrence patterns originated by the complex traffic dynamics observable during the occurrence of "noisy" network anomaly phenomena (characterized by measurable variations in the statistical properties of the traffic time series), and hence for developing qualitative and quantitative observations that can be reliably used in detecting such events.

1. Introduction

Anomaly detection is becoming an increasingly vital component of any network security infrastructure. Network anomalies, circumstances when network behavior deviates from its normal operational baseline, can arise due to various causes such as malfunctioning network devices, network overload, malicious activity, denial of service (DoS) attacks and network intrusions that disrupt the normal delivery of network services. Detecting, identifying and classifying these operational hazards is very important but at the same time hard. First of all, the definition of normal network behavior depends on several factors relating to the day to day business operations. These include traffic volume, applications running on the network, and the data they process. The goal of anomaly detection is to devise techniques that will model what a normal working network should look like and report any deviations from that definition. These techniques are typically based on machine learning, data mining or statistical analysis. The ultimate aim of anomaly detection systems is to achieve an adaptive behavior that responds in "real-time", so that problematical events can be countered as quickly as possible. However, normal behavior can only be determined by learning about past events: trends take time to learn and analyze. This paradox can only be resolved by modeling future behavior, based on a statistical idealization of the past and an observation of the present (like weather forecasting) and by specifically analyzing and observing some particularly

* *Corresponding author*. Fax: +39 081 676628.
E-mail addresses: fpalmier@unina.it (F. Palmieri), ufiore@unina.it (U. Fiore).
0167-4048/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2010.05.002

discriminating statistical features and evolutive phenomena on the network traffic. A timely response requires rapid processing of observations, and those typically originate from network monitoring devices collecting data at high rates. Consequently, designing an effective anomaly detection system involves extracting relevant information from a voluminous amount of noisy, high-dimensional data. Early integrated approach to the wider theme of *intrusion detection* were based on the assumption that most anomalous events can be revealed from the occurrence of a set of signatures flagging tampering activities and specific communication patterns in the stream of network packets. Unfortunately, while very efficient in real-time response, such systems are clueless when exposed to novel attacks, or even slight modifications of already known ones where the attack pattern does not match stored signatures or known communication behavior. On the other side, the anomaly detectors that we see on today networks are very simplistic: they observe several transport layer statistics such as the ratio of the bytes sent in each direction, the average size and mean inter-arrival time of the packets, and the results are compared against constant preassigned threshold values that are often independent from the current network utilization and the number of users. Instances where the threshold is crossed are flagged as anomalous. The trouble with existing anomaly detectors is that they take a rather narrow view of what an anomaly means, that is, they rely only on network traffic observations over a short time scale, to "explain away" observed anomalies. More specifically, the basic observation that makes traditional time-series analysis unreliable for anomaly detection is that there is an inhomogeneous pattern to human/computer resource usage, and this is clearly reflected in network resource usage. Network traffic features and hence the characteristics of probability distributions of their IP-layer packets change dynamically in the time domain (Tretyakov et al., 1998; Takayasu et al., 2000) in relation to specific network conditions. It follows that, since power laws apply to changes in traffic density, the traffic statistical characteristics change with the phase transition patterns and that their fractal-like behaviors can be affected by the packet density and its time-variation trend (Masugi and Takuma, 2007). Such a complex dynamic traffic system possesses several, almost hidden, regularities and specific statistic features that are not influenced by noise or events observable on a short time scale and hence can be more effective for investigation. Common ground between such features resides in their shared recurrence properties. That is, within the dynamical signals expressed by their associated time series we can find several stretches, short or long, of repeating patterns that are likely to be related to some hidden non-linear system properties. In this paper, we propose a novel anomaly detection scheme, particularly suitable for IP networks, based on nonlinear analysis and, more precisely, on the evaluation of the hidden recurrences and emerging non-stationary transition patterns in end-to-end traffic time series. Our scheme is based on recurrence quantification analysis (RQA) (Zbilut, 1994; Marwan and Kurths, 2002) that has been used to observe and study the above non-stationary properties in such transition patterns. Such properties, and more precisely those associated to the most informative RQA descriptors, have been used as inputs to a machine learning process, driven by support vector machines (SVM) to classify anomalous events. We demonstrated the possibility of a pure operational use of concepts and techniques derived by complex systems dynamics for developing deterministic qualitative and quantitative observations that can be reliably used in detecting anomalous events characterized by measurable variations in the statistical properties of the traffic time series. Our approach differentiates from the majority of the volume-based anomaly detection schemes by its peculiar theoretical perspective: we chose a method that does not make any specific assumption on the mathematical structure of data, does not rely on assumptions of stationarity and does not need to consider the studied traffic data as the output of a linear dynamical system.

2. Related work

Anomaly detection has been studied widely and has received an increasing attention in the last years. Most of the works in the recent research literature treat anomalies as deviations in the overall traffic volume and employ several statistical techniques for detection: exponential smoothing and Holt-Winters forecasting (Brutlag, 2000), adaptive thresholding, cumulative sum (Siris and Papagalou, 2004a; Blazek et al., 2001), maximum entropy estimation (Gu et al., 2005), and principal component analysis (Lakhina et al., 2004). Some of these works analyze the volume of aggregate traffic on a network link, others identify different flows carried on several links, and finally others look at the time series of specific kinds of packets inside aggregate traffic, restricting their focus to few kinds of attacks. The SPADE (Spade and Defense,), ADAM (Barbar et al., 2001) and NIDES (Anderson et al., 1995) systems learn a statistical frequency-based model of normal network traffic based on the distribution of most anomalous attributes like addresses and ports per transaction, and flag deviations from this model. In contrast, other anomaly detection systems like PHAD (Mahoney and Chan, 2001), ALAD (Mahoney and Chan, 2002a), LEARD (Mahoney and Chan, 2002b) and NETAD (Mahoney, 2003) monitor a larger set of fields of the packet header and use more complex time-based models, in which the probability of an event depends on the time elapsed since its last occurrence. In (Talpade et al., 1999) is presented an anomaly detection framework based on path changing, packet delay and statistic inference on the packet header information called NOMAD. The (Siris and Papagalou, 2004b) approach considered the tradeoff between the attack detection probability, the false alarm ratio, and the detection delay to estimate whether SYN-flooding attacks happened or not. With this method, the anomaly can be detected after anomalous behaviors have threatened the network, and hence the real-time detection cannot be guaranteed. More recent works have extended the range of techniques used: state-based transition analysis, neural networks, fuzzy logic, genetic algorithms, and N-gram analysis (Siris and Papagalou, 2004b; Garcia-Teodoro et al., 2009). In (Dainotti et al., 2006) is presented an automated system to detect volume-based anomalies caused by DoS attacks, combining some traditional approaches such as