Available online at www.sciencedirect.com
SCIENCE DIRECT
Computer Communications 29 (2006) 1627–1638
computer communications
www.elsevier.com/locate/comcom

Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies

Jun Jiang *, Symeon Papavassiliou [1]

New Jersey Center for Wireless Networking and Internet Security, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA

Available online 15 September 2005

Abstract

In this paper, we propose, study and analyze a new network traffic prediction methodology, based on the 'frequency domain' traffic analysis and filtering, with the objective of enhancing the network anomaly detection capabilities. Based on this approach, the traffic can be effectively separated into a baseline component, that includes most of the low frequency traffic and presents low burstiness, and the short-term traffic that includes the most dynamic part. The baseline traffic is a mean non-stationary periodic time series, and the Extended Resource-Allocating Network (ERAN) methodology is used for its accurate prediction. The short-term traffic is shown to be a time-dependent series, and the Autoregressive Moving Average (ARMA) model is proposed to be used for the accurate prediction of this component. Furthermore, it is demonstrated that the proposed enhanced traffic prediction strategy can be combined with the use of dynamic thresholds and adaptive anomaly violation conditions, in order to improve the network anomaly detection effectiveness.

Keywords: Network monitoring; Traffic prediction; Anomaly detection; Traffic analysis; Data filtering.

1. Introduction

Today's networks have become highly heterogeneous and vary greatly in the services they offer and the traffic they carry. While network heterogeneity provides more flexibility in utilizing the latest technologies and allows for customization by user applications, it also increases the risk of the occurrence of network anomalies [1,2].

With the increasing complexity of the networks, the task of detecting and preventing the network abuse becomes more and more difficult. Therefore, additional countermeasures are required. One such technique is anomaly detection. Anomaly detection uses or identifies normal behavior patterns and profiles (i.e. the expected behavior), in order to identify any unacceptable and significant deviation from the usual behavior, as possibly the result of an attack [3]. Some of the advantages of this technique is that in principle it is not restricted to any specific environment, and that can provide a way of detecting unknown attacks. The anomaly detection performance is directly correlated with the traffic prediction accuracy [4], and, therefore, accurate traffic prediction methodologies are critical for the operational effectiveness of the anomaly detection techniques.

1.1. Challenges and paper objective

The task of developing efficient and effective network traffic prediction and monitoring methodologies is mainly complicated by the following factors: (a) Users may slowly change their behavior with the system and time evolution (e.g. the traffic in a network may present changes and variations), and therefore, any associated algorithm should be capable of dynamically adapting to these changes and evolutions; (b) The analysis and prediction about the expected traffic should be based on normal data, while all the data that may be noise/anomalies and may affect the accuracy and correctness of the normal behavior prediction should be excluded; (c) In current communication networks, traffic has shown the self-similarity features and characteristics of black and brown noise. The non-stationary signals originate from the complex dynamic behavior of underlying dynamic systems, which should be best described in terms of some non-linear differential equation. However, in most cases, the particular differential equation is either

* Corresponding author. Address: 211 sussex street, 2nd FL, Harrison 07029, NJ, USA.
E-mail addresses: jun.jiang@njit.edu (J. Jiang), papavassiliou@mail.ntua.gr (S. Papavassiliou).
[1] Present address: Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece.

0140-3664/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2005.07.030

completely unknown or very difficult to estimate. Therefore, the accurate network traffic prediction is further complicated by the network noise and the occurrence of unidentified anomalous traffic. Within the scope of this paper, in the remaining we use the terms attacks and anomalies interchangeably.

This paper emphasizes on the design and development of enhanced strategies that can be used to improve the accuracy of the prediction of the network traffic normality, and as a result of the overall anomaly detection methodology, especially in cases where high burstiness is present. We first propose a methodology that provides effective traffic separation based on 'frequency domain' data analysis and filtering. Our approach is based on the observation that the various network traffic components, are better identified, represented and isolated in the frequency domain. Specifically, we separate the traffic into two main components: the baseline component and the short-term component. The baseline component includes most of the low frequency and non-stationary traffic and presents low burstiness, while the short-term component includes the most dynamic part. Since the baseline traffic is a mean non-stationary periodic time series, the ERAN algorithm [4], can be used for the accurate prediction of this part of the traffic. The short-term traffic is shown to be a time-dependent series, and the Autoregressive Moving Average (ARMA) model is proposed to be used for the accurate prediction of this component. One of the key principles of our proposed methodology is that most of the non-stationary traffic is separated and included into the baseline component, and therefore, the short-term traffic, as shown in this paper, can be well and accurately modeled using the ARMA model. Numerical results presented in this paper demonstrate that the proposed methodology of separating the traffic based on the 'frequency domain' analysis and predicting each component separately by the appropriate method, improves significantly the prediction accuracy of the total combined Internet traffic, which in turn improves the performance of the network anomaly detection as well.

The remaining of the paper is structured as follows. First, in Section 2, we present some background information and observations that motivated the proposed approach and methodology, and we summarize some related work. Section 3 describes in detail the proposed traffic separation methodology, based on the 'frequency domain' analysis, while in Section 4, specific prediction algorithms are introduced and described, that can be applied in order to predict with high accuracy the separated traffic components that resulted from the analysis and separation strategy of Section 3. In Section 5, the operational efficiency and accuracy of these methodologies in predicting the network traffic, as well as the network anomaly detection effectiveness and performance, are evaluated. Finally, Section 6 concludes the paper.

2. Background information

In order to gain some insight about the various characteristics of the Internet traffic, and present the observations that motivated our proposed strategies based on the traffic separation in the 'frequency domain', we first provide some examples regarding typical Internet traffic patterns from one dataset (Dataset 1) of real traffic collected by The Internet Traffic Archive [5], and from simulated traffic that was generated using the Optimized Network Engineering Tool (OPNET) Modeling and Simulation Tool.

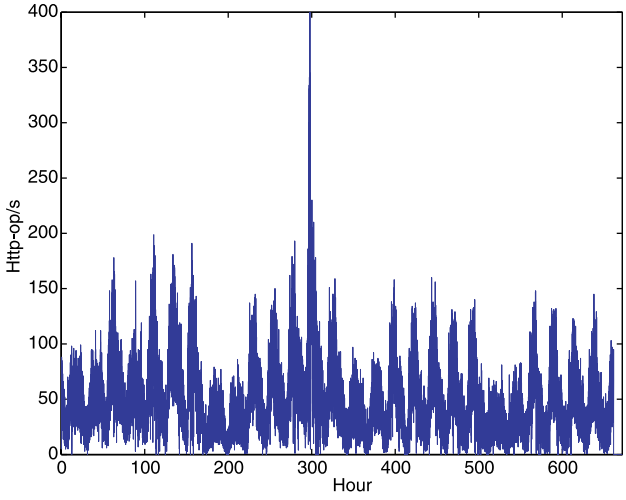

Fig. 1. Traffic Intensity on a HTTP server in Dataset 1.

Dataset 1 contains one-month's worth of all HTTP operations to the NASA Kennedy Space Center WWW server in Florida (Fig. 1). In Fig. 2, we plot the corresponding HTTP operations' intensity. We can observe from this figure that there are some underlying trends with respect to the daily and weekly traffic patterns. In Fig. 2 we present the corresponding Power Spectral Density (PSD) of the traffic in the spectral domain (frequency unit is cycles/week). We can easily observe from this figure that the daily and weekly spectrums are very strong, while most of the energy is located in the low frequency area. Similar observation can be drawn from Figs. 3 and 4, that present the corresponding data for a 4-week OPNET simulated network traffic. For the simulated traffic, we attempted to mainly represent both the non-stationary and burst characteristics of the real Internet traffic. Therefore, although there may be some small discrepancies mainly in the corresponding magnitudes, both the real traffic dataset and the simulated traffic present similar trends and behavior. It should be noted that the major identifiable traffic patterns are located within specific frequencies in both these datasets.

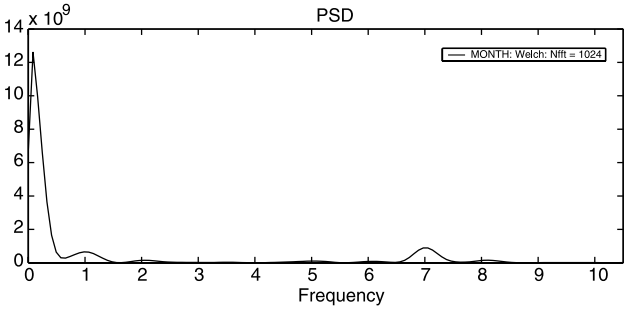

Fig. 2. Power spectral density for dataset 1(with frequency cycles/week).