Computer Networks 60 (2014) 187–200

Contents lists available at ScienceDirect

Computer Networks

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

Anomaly detection in diurnal data

Felipe Mata [a,*], Piotr Żuraniewski [b,c,d], Michel Mandjes [b], Marco Mellia [e]

[a] High Performance Computing and Networking Group, Universidad Autónoma de Madrid, Spain
[b] Korteweg-de Vries Instituut voor Wiskunde, University of Amsterdam, The Netherlands
[c] TNO, Delft, The Netherlands
[d] AGH University of Science and Technology, Kraków, Poland
[e] Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy

ARTICLE INFO

Article history:
Received 14 June 2012
Received in revised form 18 November 2013
Accepted 19 November 2013
Available online 27 November 2013

Keywords:
Anomaly detection
Diurnal pattern
Detrending
Changepoint
VoIP

ABSTRACT

In this paper we present methodological advances in anomaly detection tailored to discover abnormal traffic patterns under the presence of seasonal trends in data. In our setup we impose specific assumptions on the traffic type and nature; our study features VoIP call counts, for which several traces of real data has been used in this study, but the methodology can be applied to any data following, at least roughly, a non-homogeneous Poisson process (think of highly aggregated traffic flows). A performance study of the proposed methods, covering situations in which the assumptions are fulfilled as well as violated, shows good results in great generality. Finally, a real data example is included showing how the system could be implemented in practice.

1. Introduction

Network operators and service providers have taken a keen interest in managing the Quality of Service (QoS), and how it is perceived by their end-users (Quality of Experience). In this light, a broad range of techniques have been proposed to detect QoS degradation, see e.g. [1]. Some of these specifically focus on the Voice over Internet Protocol (VoIP) service [2], where performance degradation (due to packet loss, and increased delay/jitter) occurs during periods with high loads. Consequently, timely detection of such overload periods is crucial for management of VoIP services [3], as they enable a better cost control if applied in an automated fashion [4]. Such automated techniques rely on the statistical analysis of network traffic measurements, which commonly assumes stationarity of the data. A complication, however, is that network traffic measurements can usually *not* be considered as stationary, but rather exhibit a, roughly periodic, diurnal (day-night) pattern.

The violation of the stationarity assumption may lead to erroneous conclusions [5], in terms of large amounts of false positives/negatives. To remedy this, we propose in this paper (which builds on the results of [6]) a simple, yet effective methodology for removing the inherent daily pattern; in our study VoIP call counts data serves as the leading example. The methodology relies on the fact that the call arrival process is time-varying Poisson, which we show to be valid for the data of our case study. After removing the daily trend, we obtain standardized samples (i.e., zero mean and unit variance) that are nearly Normally distributed, as long as there is sufficient traffic aggregation—as a consequence, the fit improves when the night periods are removed from the sample (in which the chances of overload are negligible anyway). The (nearly Normal) output samples are *not* (by approximation) independent, though, which is problematic as this is required in many detection algorithms. To mitigate this effect, we propose an alternative measurement methodology that reduces the correlation for an important class of call holding

* Corresponding author. Tel.: +34 91 4972291; fax: +34 91 4972235.
E-mail address: felipe.mata@uam.es (F. Mata).
URL: http://www.hpcn.es (F. Mata).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bjp.2013.11.011

time distributions. Specifically, we show that when the call holding time distribution follows a Pareto or Log-Normal distribution, our alternative measurement procedure tends to outperform the traditional approach; this is also the case for mixtures of two Log-Normals and a Pareto distribution, which is actually the best fitting model for our VoIP data. To assess the efficacy of the resulting procedure, we have modified the overload detection methodology presented in [3] to work with Normally distributed input data and extensively tested its performance, including situations in which the independence assumption is violated; these tests convincingly show that our approach works well in great generality.

The rest of the paper is organized as follows: Section 2 presents related work. A description of the dataset is presented in Section 3. After describing how to remove the diurnal trend from the VoIP call count data in Section 4, we present the alternative measurement technique in Section 5. Next, we provide in Section 6 a description and performance evaluation of the overload detection methodology. Finally, Section 7 concludes the paper.

2. Related work

The analysis of traffic in communication networks has attracted much attention; see e.g. [7]; also its evolution in time has been studied [8]. The ubiquitous daily pattern evidently depends on the kind of users that access the network, although it can be deemed as roughly invariant (having a similar shape from day to day, that is) [4].

These users can be divided into two main groups: enterprise users, who access the network in their workplaces, and domestic users, accessing the network from their residences. The enterprise users' daily pattern is directly related to the office working hours, i.e., the load is larger during working hours, and usually there appear two clearly distinguishable peaks—before and after lunchtime (see [4] for a study on the Spanish Academic Network RedIRIS). The domestic users' pattern is also influenced by the working hours, but obviously in the opposite way: the load is larger after the usual working hours (see [9] for such a study held within the European project TRAMMS).

This shape invariance is actually observed at different timescales. For instance, on a weekly basis, the shape of the pattern is approximately the same from Monday up to Thursday. On Fridays, we observe a scaled version of the other working days pattern—i.e., the shape is essentially the same, but the load is somewhat smaller. Finally, on weekends and holidays, we found almost flat patterns when dealing with enterprise measurements—principally due to applications that are left running and generate traffic without user interaction.

A similar conclusion holds when focusing on VoIP traffic only, see e.g. [8,10]. Here it is noted that these VoIP-related studies primarily focus on call characteristics (in terms of the call arrival process and call holding time distribution) rather than daily/weekly patterns. The call arrival process is widely accepted to be accurately modeled by a time-inhomogeneous Poisson process (roughly stationary at short timescales, ranging from minutes to hours [8,11]). Conversely, there is no consensus as to which model should be used for the call holding times (where it *is* clear that the exponential distribution is *not* a good candidate). A broad range of distributions have been proposed, such as the hyper-exponential [8], the inverse Gaussian [10], and the Log-Normal [12]. The trend-removal issue can be approached relying on general traffic forecasting techniques [13], or by time series with seasonal cycles [14].

3. Dataset description

The experiments reported on in this paper are using actual traffic traces collected from an operational network. Using Tstat [15], we monitored IP traffic exchanged by customers in a large Point of Presence (POP) of an operator in Italy where VoIP is deployed. A total of about 22,000 customers were continuously monitored for more than 4 months, starting from November 2010. Tstat was used to identify VoIP flows, i.e., voice calls, and to extract several performance indexes for each call [10]. In particular, in the context of the present paper we are interested in the call arrival process and call holding time distribution. The resulting dataset contains the log of the call arrival epochs and the corresponding durations. Later in this paper, we statistically analyze these, and use the resulting processes/distributions to assess the performance of our algorithm. The dataset containing start and end times of the calls will be referred to as *detailed* below.

However, storing detailed call records is not practical for history analysis in large-scale networks. In these cases, summarized statistics are stored instead, and detailed logs are kept for short time periods (say several days). Anomaly detection is performed in the summarized datasets, and the detailed records are used for further forensic analysis on the relevant events encountered. Consequently, we build a *summarized* dataset from the detailed logs for its application in the trend removal methodology described in Section 4. To construct the summarized dataset, we adopt the traditional way that records the number of calls present in the system at equidistant points in time (e.g., $N_0, N_t, N_{2t}, \ldots$). The separation t between records is set to 5 minutes for this dataset, based on the results presented in the following subsection.

3.1. Call arrival process

The Poisson process is the classical model for the arrival process of voice calls. Evidently, at longer timescales this model does not match with reality, due to the absence of a day-night pattern (and a weekly pattern). To cope with this effect, non-homogeneous Poisson processes are used instead, where the arrival rate is usually assumed constant for blocks of time, of say, L minutes. To verify the 'local Poisson claim' for some L, we apply to our detailed dataset a test presented in [11]. To construct the test, we split up a day into disjoint blocks of length L, resulting in a total of I blocks. Let T_{ij} be the jth arrival time in the ith block. Denoting with J_i the total number of arrivals within the ith block, we then define $T_{i0} = 0$ and for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$,