Anomaly Detection in VoIP Traffic with Trends

Felipe Mata*, Piotr Zuraniewski†§¶, Michel Mandjes† and Marco Mellia‡
*High Performance Computing and Networking Group, Universidad Autónoma de Madrid, Spain
†Korteweg-de Vries Instituut voor wiskunde, University of Amsterdam, The Netherlands
‡Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino
§TNO, Delft, The Netherlands
¶AGH University of Science and Technology, Kraków, Poland

***Abstract*—In this paper we present methodological advances in anomaly detection, which, among other purposes, can be used to discover abnormal traffic patterns under the presence of deterministic trends in data, given that specific assumptions about the traffic type and nature are met. A performance study of the proposed methods, both if these assumptions are fulfilled and violated, shows good results in great generality. Our study features VoIP call counts, but the methodology can be applied to any data following, at least roughly, a non-homogeneous Poisson process (think of highly aggregated traffic flows).**

I. 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, 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 [2], as they enable a better cost control if applied in an automated fashion [3]. 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 [4], in terms of large amounts of false positives/negatives. To remedy this, we propose in this paper 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 time distributions; for our VoIP data we show that the best fitting model for the call holding times is a mixture of two log-normals and a Pareto distribution, which is included in this class. To assess the efficacy of the resulting procedure, we have modified the overload detection methodology presented in [2] 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 II presents related work. A description of the dataset is presented in Section III. After describing how to remove the diurnal trend from the VoIP call count data in Section IV, we present the alternative measurement technique in Section V. Next, we provide in Section VI a description and performance evaluation of the overload detection methodology. Finally, Section VII concludes the paper.

II. Related Work

The analysis of traffic in communication networks has always attracted much attention; see e.g. [5]; even its evolution throughout time has been studied [6]. 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) when the kind and number of users are fixed [7]. A similar conclusion holds when focusing on VoIP traffic only [8], [6], where 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 [6], [9]). 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 [6], the inverse Gaussian [8], and the log-normal [10]. The trend-removal issue can be approached relying on general traffic forecasting techniques [11], or by time series with seasonal cycles [12].

III. Dataset Description

Experiments in this paper are using actual traffic traces collected from an operational network. Using Tstat [13], 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 22,000 customers were continuously monitored for more than 4 months, starting from November 2010. Tstat is used to identify VoIP flows, i.e., voice calls, and to extract several performance indexes for each call [8]. 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.

A. 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. As our data set consists of 5 minutes samples, we wish to verify the 'local Poisson claim' for some L being a multiple of 5 minutes. To this end, we apply to our detailed dataset a test presented in [9]. 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 j^{th} arrival time in the i^{th} block. Denoting with J_i the total number of arrivals within the i^{th} block, we then define $T_{i0} = 0$ and for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$,

$$R_{ij} := (J(i) + 1 - j)\Big(- \log\Big(\frac{L - T_{ij}}{L - T_{i,j-1}}\Big)\Big). \qquad (1)$$

Under the null hypothesis (arrival rate is constant within each block), the R_{ij} are independent standard exponential variables; see [9] for further background and a justification of the test.

In Table I we present the results of applying the test to different block sizes L. We use the Kolmogorov-Smirnov (KS) test to verify the null hypothesis at the 5% significance level. The results presented in the table indicate that the arrival process can be regarded as non-homogeneous Poisson at relatively short timescales only, say less than 10 minutes (which is sufficient for our objectives later on in this paper).

B. Call Holding Time (CHT) Distribution

In the literature it is generally concluded that the CHT is poorly modeled by the exponential distribution. Instead, distribution with heavier tails have been proposed. A visual inspection using *log-log plots* of the empirical Complementary CDF (CCDF) of the sample, which allow us to gain insight in the tail of the distribution, evidences the heavy-tailed nature in our dataset as well. Consequently, in our Goodness-of-Fit (GoF) assessment we restrict ourselves to heavy-tailed distributions (that is, heavier than the exponential distribution).

TABLE I
Results of the arrival process assessment.

L (min)	Rejection %	L (min)	Rejection %
90	74%	25	27%
60	61%	20	19%
45	50%	15	14%
35	39%	10	9%
30	35%	5	7%

TABLE II
GoF results for different fitting models.

Distrib.	Parameters			KS statistic
Pareto + 2 Log-Ns	Pareto p = 0.6793 k = 0.2749 σ = 63.1607	Log-N$_1$ p = 0.2023 μ = 6.0857 σ = 0.9523	Log-N$_2$ p = 0.1184 μ = 3.5410 σ = 0.5201	0.0046
2 Log-N	Log-N$_1$ p = 0.1089 μ = 3.6421 σ = 0.4810	Log-N$_2$ p = 0.8911 μ = 4.2926 σ = 1.5528	-	0.0074
Weibull + Log-N	Weibull p = 0.0968 λ = 42.7199 k = 2.4978	Log-N p = 0.9032 μ = 4.2964 σ = 1.5385	-	0.0075

To measure the GoF we use again the KS test. Such a procedure is in principle not justified, as we are *estimating* the parameters of the hypothesized models from the sample [14]. However, we still can use the KS statistic as a measure of model discrepancy, so as to select the best fitting model.

Table II presents the models that fitted the data reasonably well. We sorted them by the value of the KS statistic (the lower the better), along with the Maximum Likelihood Estimates (MLE) of the corresponding parameters. Fig. 1 shows the *log-log* plot of the empirical CCDF of the data along with the models presented in Table II. In line with the quantitative results, we observe in the figure that the best fit is provided by the mixture of two log-normal and a Pareto distribution. This mixture model is capable of fitting the whole body of the data, while there is a slight lack of fit in the very end of the tail. Furthermore, we can observe a small oscillation in the tail corresponding to the data, not captured by any of the fitting models, probably due to the lack of samples of this size.

IV. Detrending Methodology

A. Methodology Description and Expected Results

Our methodology exploits the presence of a weekly pattern to estimate and remove the seasonality from the measurements, obtaining a sequence of zero-mean, unit-variance observations. In our set-up the measurements are time series of traffic metrics (think of number of active calls) at a given time granularity — 5 minutes in our study. These measurements

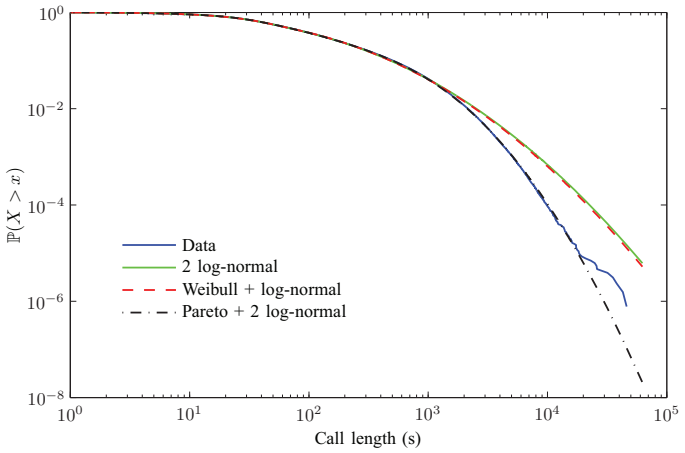


Fig. 1. CCDF *log-log* plots of the data and the best fitting models according to the KS statistic value.

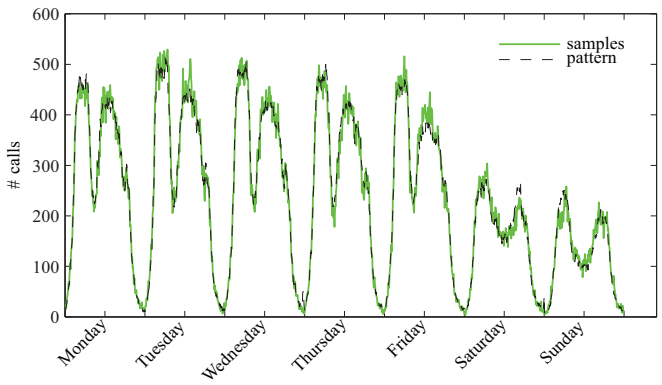


Fig. 2. Data samples for the week under study and estimated pattern based on previous weeks data samples.

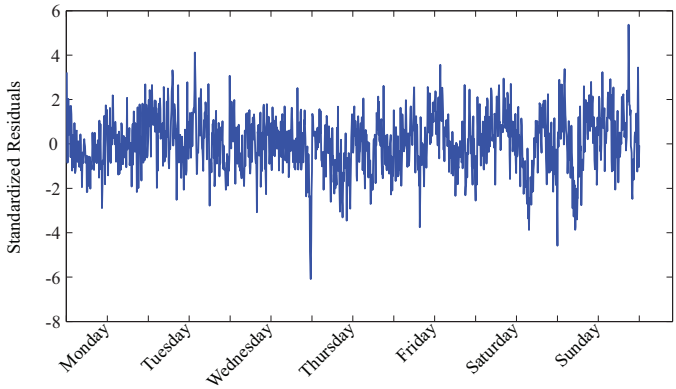


Fig. 3. Residuals obtained after standardization with the estimated pattern.

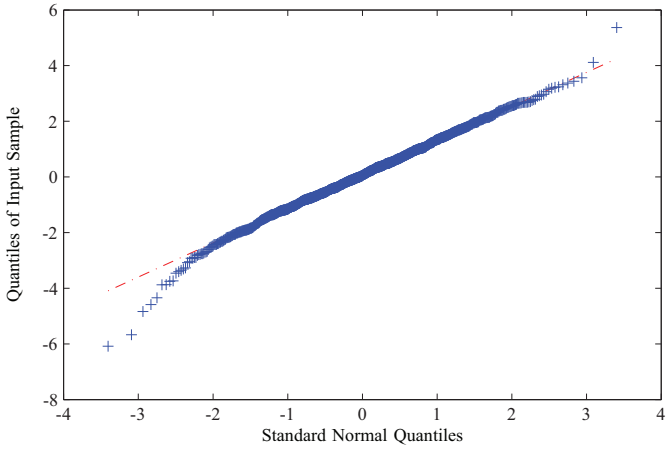


Fig. 4. Gaussian Quantile-Quantile plot of the residuals.

are denoted by x_i^n, with $i = 0, 1, 2, \ldots, 2015$ corresponding to the 5-minute interval within the week (starting on Monday midnight), and n to the week number within the dataset (out of a total of N weeks). The goal is to provide a good estimate $\boldsymbol{y}^n$ for the measurement vector of week n, $\boldsymbol{x}^n$, using the information available from previous weeks, $\boldsymbol{x}^j$, $j < n$, assuming the differences from week to week in the weekly pattern to be due to random deviations from an average network usage pattern. To this end, we propose the following model for the measurements: $\boldsymbol{x}^n = \boldsymbol{\alpha} + \boldsymbol{\varepsilon}^n$, where $\boldsymbol{\alpha}$ denotes the average pattern and $\boldsymbol{\varepsilon}^n$ are the random deviations from such pattern.

The proposed estimation computes the trend $\boldsymbol{y}$ as an arithmetic average of the observations of the last $w = 5$ weeks. This window size w balances accuracy and robustness to pattern shifts quite well; we have also tested different averaging processes, but the differences in performance are negligible.

We can then remove the estimated pattern from the actual measurements, so as to obtain zero mean residuals. Recalling that our arrival process is locally Poisson (cf. Section III-A), and because Poisson variates with a high mean are approximated well by the Normal distribution, the resulting residuals (by approximation) form a sequence of zero mean Normally distributed random variables. Note, however, that they are not homoscedastic (that is, they do *not* have a uniform variance). Therefore, they need to be standardized, which can be done by dividing each residual by its standard deviation. Instead of designing another model for estimating the pattern variance, we exploit the fact that for Poisson random variables, the mean and the variance are equal. Hence, we obtain standardized residuals through

$$\boldsymbol{r}^n = \frac{\boldsymbol{x}^n - \boldsymbol{y}^n}{\sqrt{\boldsymbol{y}^n}}. \tag{2}$$

B. Model Performance Results

We have computed the (standardized) residuals in our dataset using (2). For the sake of brevity, we only show the results for one week, which turned out to be highly representative. We found that the variance of the residuals was substantially higher than expected, with some values exceeding 5σ. This is caused by the fact that during nights the load decreases drastically, thus leading to large residuals (due to the small numerator in (2)). For the further analysis, we have then decided to remove the nights (defined as periods from midnight to 6 a.m.), which is justified as the chances of overload during the nights are negligible anyway.

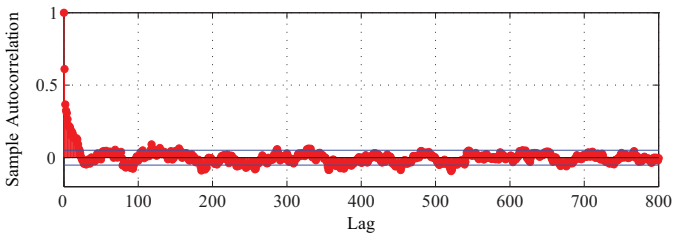


Fig. 5. Autocorrelations of the residuals (nights removed).

The corresponding estimated pattern and residuals without night periods are shown in Fig. 2 and Fig. 3, respectively. Compared to their counterparts with night included (not shown here due to space limitations), we have observed that leaving out the nights reduces the variance substantially. Also Fig. 4, which shows the Gaussian Q-Q plot of the residuals without nights, indicates that Normality holds. We present also in Fig. 5 the autocorrelations of the residuals. This graph shows that the residuals are *not* independent — note that the horizontal lines indicating the 95% confidence interval around 0 are exceeded, particularly for the first lags; we return to this issue later in great detail. In addition, there is a periodic component in the residuals. As this periodicity, which is likely to be due to the inherently simple nature of the detrending procedure, is relatively weak, we do not take it into account in our study.

Statistical overload detection procedures usually assume that the observations used are independent. As a result, when using our residuals in such a test, the high correlation may lead to a significant performance degradation (in terms of high numbers of false positives and negatives). We are inclined to think that the correlation in the residuals is essentially due to the simplicity of our trend estimation model, which is not able to adapt dynamically — at a short timescale, say hours — to deviations from the pattern. As a consequence, when actual measurements are above the estimated pattern from previous weeks, there is a high probability that this situation would remain the same for the next samples, and vice versa. Nonetheless, we have decided to keep the model as simple as possible at the expense of modest performance degradations (which can be controlled as described in Section VI).

V. Measurement Alternative

As mentioned above, the correlations in the residuals of our detrending procedure are likely to be due to the way the measurements are obtained. Traditionally, detection procedures record the number of calls present in the system at equidistant points in time (e.g., $N_0, N_t, N_{2t}, \ldots$). In this section we analyze an alternative, and compare its performance (in terms of correlation) with the traditional one. More precisely, we define M_a as the number calls present during the interval $[a, a+t]$, for a given t.

A. Alternative Procedure

To evaluate the performance of this alternative, we compute the correlation between two measurements at different time instants — e.g., M_0 and M_{kt}. To simplify the following computations, we assume that the arrival process is Poisson with constant arrival rate. Consequently, we obtain

$$\mathbb{C}\mathrm{orr}(M_0, M_{kt}) = \frac{\mathbb{C}\mathrm{ov}(M_0, M_{kt})}{\mathbb{V}\mathrm{ar}(M_0)} = \frac{\mathbb{C}\mathrm{ov}(M_0, M_{kt})}{\mathbb{E}[M_0]}, \quad (3)$$

using the Poissonian and stationarity assumptions. To compute (3), we define
$A_a = \{\#$ arrivals up to time a that depart in $[a, a+t]\}$
$B_a = \{\#$ arrivals in $[a, a+t]$ that depart in $[a, a+t]\}$
$C_a = \{\#$ arrivals in $[a, a+t]$ that are still present at $a+t\}$
$D_a = \{\#$ arrivals up to time a that are still present at $a+t\}$,
obtaining the following identity:

$$M_a = A_a + B_a + C_a + D_a. \quad (4)$$

Now notice that calls that are there at the end of the first interval are potentially still present at the beginning of the second interval, so that we only have to take into account C_0 and D_0. Also, only calls that arrived before the beginning of the second interval can interact, so for the same reason we only need to include A_{kt} and D_{kt}. Consequently, (3) can be rewritten as follows:

$$\mathbb{C}\mathrm{orr}(M_0, M_{kt}) = \frac{\mathbb{C}\mathrm{ov}(C_0 + D_0, A_{kt} + D_{kt})}{\mathbb{E}[A_0 + B_0 + C_0 + D_0]} \quad (5)$$

Also realize that for all $a \in \mathbb{R}$: $C_a + D_a = N_{a+t}$ and $A_a + D_a = N_a$. Therefore, the numerator in (5) is

$$\mathbb{C}\mathrm{ov}(N_t, N_{kt}) = \rho \mathbb{P}(S_e > (k-1)t), \quad (6)$$

where in the last identity we have defined $\rho = \lambda \mathbb{E}[S]$, S being the service time distribution (with finite mean), and S_e its excess lifetime; the equality in (6) is well-known from queueing theory. It is also true that $B_a + C_a = F_a$ is the number of arrivals in the interval $[a, a+t]$. This can be used to compute the mean of M_a for all $a \in \mathbb{R}$:

$$\mathbb{E}[M_a] = \mathbb{E}[N_{a+t} + F_a] = \rho + \lambda t = \rho\Big(1 + \frac{t}{\mathbb{E}[S]}\Big). \quad (7)$$

Finally, using (6) and (7), we obtain the result for (3):

$$\mathbb{C}\mathrm{orr}(M_0, M_{kt}) = \frac{\rho \mathbb{P}(S_e > (k-1)t)}{\rho(1 + \frac{t}{\mathbb{E}[S]})} = \frac{\mathbb{P}(S_e > (k-1)t)}{1 + \frac{t}{\mathbb{E}[S]}}. \quad (8)$$

It is worth noting that in the previous computations we did not make any assumption regarding the service time distribution, which means that (8) holds for any kind of service time distribution given that the arrival process is Poisson.

B. Correlations Study: Impact of Service Time Distribution

We now compare the correlation resulting from (8) with the one from the traditional call count process N_t, assuming specific service distributions; the main question is whether or not $\mathbb{C}\mathrm{orr}(M_0, M_{kt}) < \mathbb{C}\mathrm{orr}(N_0, N_{kt})$; if yes, then the correlation is reduced.

For the exponential distribution, it is easy to verify that the inequality is never satisfied, so that the traditional method is preferred for exponential service times. For the Pareto

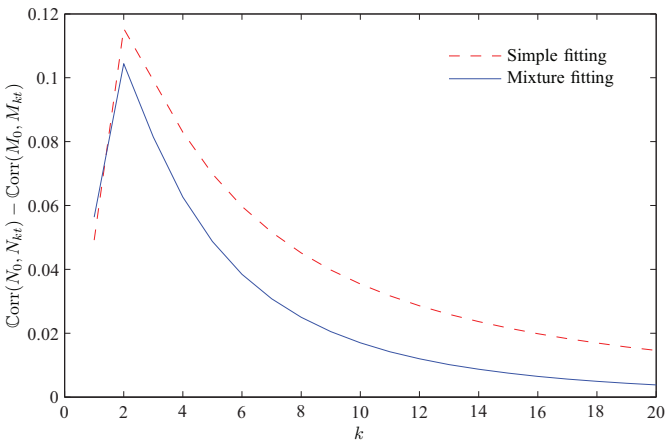


Fig. 6. Correlation comparison of both measurement alternatives assuming the call holding time is distributed accordingly to the two best fitting models as presented in Section III-B.

distribution with $s > 1$ and parameter α, the corresponding excess lifetime distribution reads

$$\mathbb{P}(S_e > y) = \frac{1}{\mathbb{E}[S]} \int_y^\infty \mathbb{P}(S > \tau) d\tau = (1+y)^{1-\alpha}. \quad (9)$$

It turns out that we need to check whether

$$f(t) := \left(1 - \frac{t}{1+kt}\right)^{1-\alpha} < 1 + (\alpha - 1)t =: g(t). \quad (10)$$

To find a sufficient solution for this new inequality, we will use the fact that, if $f(0) \leq g(0)$, then $f'(t) \leq g'(t)\ \forall t > 0$ guarantees $f(t) \leq g(t)$. Applying this argument three times, we obtain the following sufficient condition: $k > \alpha/2$. However, we observed by numerical analysis that the condition is even less restrictive: for some cases with $k < \alpha/2$ the inequality still holds, and the alternative procedure is to be preferred.

For some other distributions, like log-normal, it is not possible to obtain any analytical results; hence, the inequality needs to be verified numerically. We do not include these experiments here; instead we provide an example in which we compare both measurement alternatives, assuming the call holding time obeys the best fitting models obtained in Section III-B (a simple log-normal fit, and the mixture with one Pareto and two log-normal components). Fig. 6 shows the difference between both models, indicating that the alternative procedure outperforms the traditional one. The difference between both correlations is maximal at the first lags, which is the timescale of interest for the purposes of the methods presented in this paper. We have therefore found practical evidence using actual network measurements that the alternative measurement procedure reduces correlations, and is hence more suitable to be used in the statistical detection procedures.

VI. Overload Detection Methodology

In [2] an overload detection algorithm is studied for an M/G/∞ queuing system, relying on the testing framework of [15, Section VI.E]. It was also pointed out in [2] that the detection procedure in [15] for a changepoint in the mean of i.i.d. Normally distributed samples (with known and constant variance) can be adapted to a changepoint in the variance (with known and constant mean). (Semi-)closed-from results were included in [2], but the performance of the resulting test was not evaluated. Furthermore, the case of a *simultaneous* change in both mean and variance was not covered in [2].

Below we provide a detailed analysis featuring a procedure to detect a simultaneous change of the mean and variance. We asses its performance in case the independence assumption is fulfilled, but also when it is violated due to non-negligible correlations. The latter case is obviously highly relevant in our VoIP context: as we saw in Section IV, our detrending method leaves some autocorrelation in the residuals. If the changepoint detection method developed for the independent case works sufficiently well (at least up to some specific level of correlation), it may be an attractive and viable alternative to the idea of explicitly incorporating correlation into the test (which will lead to a considerably more complicated procedure).

A. Changepoint Detection

We wish to detect a changepoint in the Normally distributed data, that is, whether during our observation period the parameter vector (μ, σ) (which we denote as the probability model $\mathbb{P}$) changes into $(\nu, \eta) \neq (\mu, \sigma)$ (the model $\mathbb{Q}$). More formally, we consider the following (multiple) hypotheses. Let X_i be the sequence of independent observations obtained from the Normal distribution.

H_0: $(X_i)_{i=1}^n$ are distributed according to a Normal random variable with parameters vector (μ, σ).

H_1: For some $\delta \in \{1/n, 2/n, \ldots, (n-1)/n\}$, it holds that $(X_i)_{i=1}^{\lfloor n\delta \rfloor}$ is distributed according to a Normal random variable with parameters vector (μ, σ), whereas $(X_i)_{i=\lfloor n\delta \rfloor + 1}^n$ is distributed according to Normal random variable with parameter $(\nu, \eta) \neq (\mu, \sigma)$.

Following the notation used in [2], we consider the likelihood-ratio test statistic (cf. Neyman-Pearson lemma):

$$\max_{\delta \in [0,1)} \left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor + 1}^{n} L_i - \varphi(\delta) \right), \text{ with } L_i := \log \frac{\mathbb{Q}(X_i)}{\mathbb{P}(X_i)}, \quad (11)$$

for some function $\varphi(\cdot)$ which is defined shortly. If the test statistic is larger than 0, we reject H_0. The function $\varphi(\cdot)$ is introduced to get an essentially uniform alarm rate with respect to δ. As in [2], $\varphi(\cdot)$ is given in the implicit form using Legendre transform $I(u) = \sup_\vartheta(\vartheta u - \log M(\vartheta))$ of the moment generating function $M(\vartheta)$:

$$\delta I\left(\frac{\varphi(1-\delta)}{\delta}\right) = \alpha^\star, \quad (12)$$

where $\alpha^\star = -\log\alpha/n$; here α is a measure for the likelihood of false alarms (for instance 0.05) and

$$M(\vartheta) = \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{\vartheta}{2}\left(\frac{x-\nu}{\eta}\right)^2 - \frac{1-\vartheta}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\mathrm{d}x}{\sqrt{2\pi}\eta^\vartheta\sigma^{1-\vartheta}}, \tag{13}$$

so that

$$\log M(\vartheta) = -\frac{\frac{\vartheta(1-\vartheta)(\mu-\nu)^2}{2\eta^2(1-\vartheta)-2\vartheta\sigma^2} + \vartheta\log\eta + (1-\vartheta)\log\sigma}{\frac{1}{2}\log\left(\frac{\vartheta}{\eta^2}+\frac{1-\vartheta}{\sigma^2}\right)}$$

It is possible to explicitly find the value $\vartheta^\star$, optimizing $I(u) = \vartheta^\star(u)\,u - \log M(\vartheta^\star(u))$ (not shown due to space limits), thus to numerically solve Eq. (12) and to obtain the threshold function $\varphi(\cdot)$, and the test statistic (11).

B. Analysis with Synthetic Data

It is clear that applying the above test to the (approximately Normally distributed) residuals that we generated in Section IV-A, allows us to detect whether or not a given pattern shows substantially higher values than those suggested by the trend. Alternatively, one may be interested in tests that detect whether or not the offered load is getting close to the system's capacity. Below we point out how to set up a test that focuses on anomalies of the latter kind; a procedure to detect anomalies of the former kind can be set up similarly.

Due to the normalization procedure (2), in our case the parameters in model $\mathbb{P}$ are equal to $(\mu,\sigma) = (0,1)$ while these related to the model $\mathbb{Q}$ will typically be provided by the system administrator who derives them after assessing what is for instance an acceptable overload probability level. An example of such calculations, based on the Erlang model, is presented in [2, Sec. 5, Example 1]. The figures provided there were the following: an expected number of users during 'busy hour' was equal to 320 and a number of users which, if reached, was considered to be overload was equal to 375 (that was the value of the parameter one tested against if considering a counterpart of the model $\mathbb{Q}$ presented here). For the finite capacity system, with these numbers a blocking probability would be around 0.1%. If we then use these figures to calculate the parameters of the model $\mathbb{Q}$ presented here, we get due to Eq. (2),

$$\nu = \frac{375-320}{\sqrt{320}} \approx 3.075, \quad \eta = \sqrt{\frac{375}{320}} \approx 1.083 \tag{14}$$

In our situation, where the expected number of calls is not constant but fluctuates (cf. Fig. 3), to keep a (roughly) constant blocking probability, one would have to constantly update the values of (ν,η) to be tested against. While this is entirely possible, from a practical standpoint it may be more attractive to stay with just one fixed pair. For the number of calls larger than, say, 100 one would not observe much change in blocking probability despite changing the parameters values, while for lower numbers of calls a blocking probability will be somewhat underestimated.

All the changepoint detection tests presented below use parameters $(\mu,\sigma) = (0,1)$ (model $\mathbb{P}$) and $(\nu,\eta) = (3.075, 1.083)$ (model $\mathbb{Q}$) unless explicitly stated otherwise. However, we want to underline, that the algorithm for changepoint detection in Normally distributed data we provide here is generic in that sense that any values (μ,σ) and (ν,η) can be used. Moreover, we may abstract from our VoIP situation, and use the proposed method for any type of data obeying the Gaussianity assumption.

1) Synthetic Independent Data: Here we will present the results of an experiment ($\mathbf{E}_1$) in which we draw 200 independent samples from the distribution $\mathbb{P}$, followed by another 200 samples from the distribution $\mathbb{Q}$. We take a window of length 50 samples, that is, we test whether H_0 should be rejected based on data points $X_i,\ldots,X_{i+49}$, for $i = 1$ up to 351. The first window in which the influence of the parameters (ν,η) is noticeable is therefore window number 152. 5000 independent repetitions of this experiment were run to assess the performance of the method. In Fig. 7(a) we see that before the changepoint (red vertical line) the detection ratio, which is in this case an estimate of the false alarm ratio (type I error probability), is about the assumed level of 5%. Already, when only one observation from the distribution $\mathbb{Q}$ is present (window number 152 as explained above), the detection ratio (which now estimates the power of the test) is about 75%. The detection ratio then increases sharply as more samples from the new distribution appear. Furthermore, the position of the changepoint returned by the test is in general very close to the true one, as indicated on Fig. 7(b). One has to bear in mind, however, that for a less pronounced change (for instance a gradual change) the detection ratio would grow slower, cf. the results in [2].

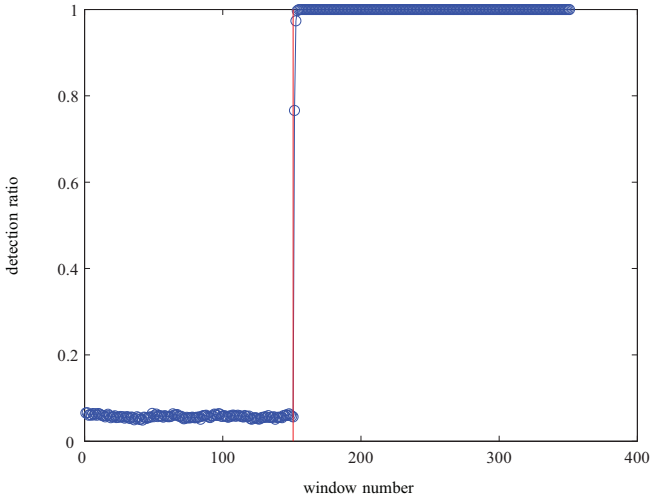
2) Synthetic Dependent Data: As it was stated earlier, the considered changepoint detection test is designed for i.i.d. data. Nevertheless, in practice one may be interested in the performance of the proposed algorithm in case of *dependent* data. Below we present the results of an experiment ($\mathbf{E}_2$), which is similar to $\mathbf{E}_1$, but now the observations before and after the changepoint originate from AR(1) processes (autoregressive processes, see for example [16]) with different levels of correlation. The AR(1) process is defined as

$$X_i - \mu = \phi\,(X_{i-1}-\mu) + \epsilon_i$$

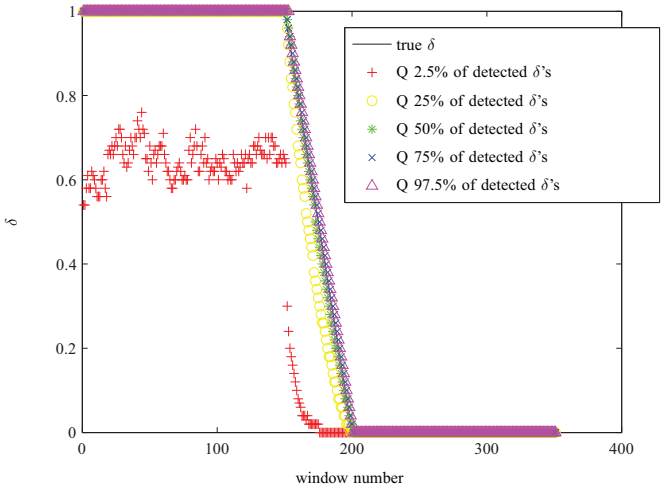
where $\{\epsilon_i\}$ is a sequence of i.i.d. random variables with zero mean and variance τ^2, has mean μ, and autocorrelation function

$$\gamma(k) = \phi^k, \text{ for } k = 0,1,\ldots$$

Note, that because of the relationship $\mathbb{V}\mathrm{a}rX_i = \tau^2/(1-\phi^2) = \sigma^2$ the value of τ used to generate the sample after the change will be equal to $\tau = 1.083\sqrt{1-\phi^2}$ (not just 1.083). The findings related to $\mathbf{E}_2$, for $\phi \in \{0.2, 0.8\}$, are presented in Fig. 8 and show how the performance of our method is degraded by increasing the level of autocorrelation. In $\mathbf{E}_3$ we consider a larger set of autocorrelations: $\phi \in \{0.2, 0.4, 0.6, 0.8\}$; the results are given in Table III.



(a)



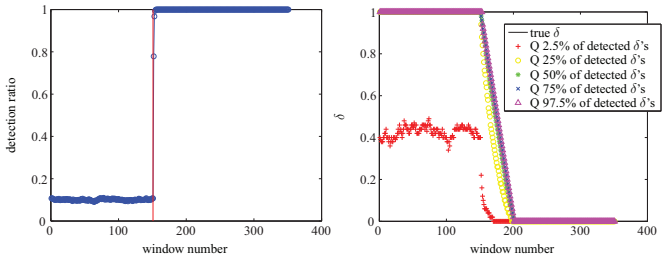
(b)

Fig. 7. $\mathbf{E}_1$: (a) detection ratio; (b) distribution of detected changepoints.

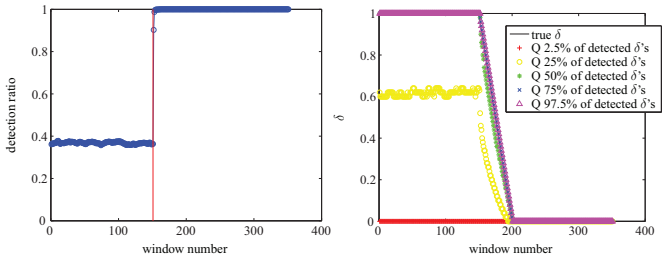
TABLE III

$\mathbf{E}_3$ AND $\mathbf{E}_4$: IMPACT OF CORRELATION ON FALSE ALARM RATIO AND DETECTION RATIO. UPPER PART: $\alpha = 5\%$ — LOWER PART $\alpha = 0.5\%$

ϕ	$\mathbf{E}_3$ mean of detection ratio in windows 1–151	$\mathbf{E}_4$ false alarm ratio	$\mathbf{E}_3$ detection ratio for window no. 152	$\mathbf{E}_4$ detection ratio for true $\delta = \frac{49}{50}$
0	5.7%	5.7%	76.6%	76.8%
0.2	10.1%	5.3%	77.9%	77.0%
0.4	17.7%	10.4%	80.8%	79.8%
0.6	27.2%	17.9%	85.9%	82.7%
0.8	36.8%	24.0%	90.3%	88.8%
0	0.6%	0.6%	45.6%	46.0%
0.2	1.8%	0.7%	47.6%	45.7%
0.4	5.4%	1.8%	49.2%	47.8%
0.6	12.7%	5.4%	54.8%	50.3%
0.8	23.8%	11.3%	59.3%	50.9%



(a) $\phi = 0.2$.



(b) $\phi = 0.8$.

Fig. 8. $\mathbf{E}_2$ for different autocorrelation levels; left column: detection ratio; right column: distribution of detected changepoints.

The detection ratios in $\mathbf{E}_2$ should be interpreted with care: in case of non-negligible correlation, the relative frequency of detections before the actual change happens is not anymore an unbiased estimator of type-I error probability. This is because if the test incorrectly detected a changepoint using data contained in k^{th} window, in the next step window $k+1$ contains the same data apart form the oldest observation (which is dropped) and the appended newest observation which is now highly dependent on several previous ones. As a result, the chance of spurious detection is increased and at the same time, also a power of the test (detection ratio) is affected. To assess this effect, we performed an additional experiment ($\mathbf{E}_4$), that is the same as $\mathbf{E}_3$, apart from the fact that samples of length 50 are generated. In other words, the size of these samples equals the detection window, thus completely 'regenerating' the input data for each of the 5000 repetitions.

For the somewhat lower value of $\phi = 0.2$ (Fig. 8) the spurious detection ratio is about 10% (versus the prescribed 5%, which is achieved in case of the 'regenereted' samples of $\mathbf{E}_4$), while the position of the detected changepoint is close to the true one. In other words, for these low values of ϕ the detection procedure performs well.

When increasing ϕ the performance degrades: the false alarm ratio increases, while also the position of the detected changepoint becomes less accurate. If the false alarm ratio is regarded to be too high, a quick fix is to lower the nominal value of α. Obviously, again the price we pay lies in the detection ratio. $\mathbf{E}_3$ and $\mathbf{E}_4$ were redone with $\alpha = 0.005$; see the lower part of Table III. One can see that for example for $\phi = 0.6$ the false alarm ratio is now close to the prescribed value. This indicates that by an appropriate tuning the test can be adjusted such that it has the desired performance.

C. Results on Real Data Trace

In this section we present results obtained by applying our anomaly detection method to a real data trace. To demonstrate

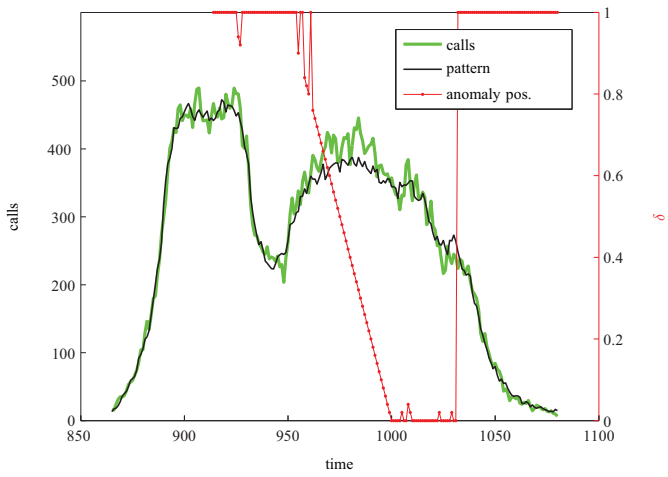


Fig. 9. Real data example

its performance, we select one day (the Friday of Week 11) from our repository and give detailed comments about the outcome of the tests. Figure 9 should be interpreted as follows.

- On the *left scale* we record the actual and average number of calls (a pattern) based on observations from five previous weeks, as indicated in Section IV-A.
- On the *right scale* we have a relative position of the anomaly detected in the window of 50 samples which ends at the given time point (meaning that we have decided to skip the first 49 values as they would require readings from a previous day). A value of 1 means 'no anomaly detected', while for example a value of 0.94 observed at time point no. $t = 926$ means that at that moment a system reports an anomaly, and declares it has happened 3 observations before ($t = 923$) (as for a detection window of size 50 distance between two consecutive observations is 0.02 and $1 - 0.94 = 0.06 = 3 \cdot 0.02$). Later on, we again observe a series of readings with no alarm reported. Then, at the onset of the 'afternoon peak' ($t = 955$), after some uncertainty at the beginning, we observe a consistent period that the detector reports that the number of calls is significantly higher than average, which is confirmed by a visual inspection. From $t = 1032$ on, the system again declares no anomaly.

Observe, that the proposed method is capable of not only detecting an overload, defined as a situation when the system approaches its capacity limits, but also the situation that the number of calls, while still being below the aforementioned capacity limits, grows (falls) faster (slower) than the trend (see also the remark at the start of Section VI-B). Such information can be useful for example in call centers, as it may indicate the need for more staff than was initially planned.

VII. Concluding Remarks

We have discussed the problem of anomaly detection in the situation that a strong trend (diurnal pattern being a result of human behavior) is present in the analyzed sample. As such a trend is a kind of nonstationarity by itself, its presence in most cases has a negative effect on the performance of any changepoint detection algorithm. Thus, the anomaly detection method we proposed in our paper consists of two steps: trend estimation and removal, which results in obtaining so-called residuals, and then applying a changepoint detection method to those residuals. Our contribution to the first step is verifying and exploiting the fact that the arrival process is non-homogeneous Poisson, which leads to a straightforward, yet effective trend removal procedure. The contribution to the second phase is twofold: a methodology to simultaneously detect changes in mean and variance, and extensive tests for the cases of both independent and correlated input. Besides, we also present and discuss a measurement procedure that leads to potentially significant correlation reduction. Finally, a real data example is included showing how the system could be implemented in practice.

As future work, we plan to further analyze the time-inhomogeneity of the CHT distribution, and which implications it may have in the proposed methodology. Furthermore, we plan to extend our work to multi-service measurements, where the Poissonian assumption typically does not hold.

References

[1] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, 2007.
[2] M. Mandjes and P. Żuraniewski, "M/G/∞ transience, and its applications to overload detection," *Perf. Eval.*, vol. 68, no. 6, pp. 507–527, 2011.
[3] F. Mata, J. Aracil, and J. L. García-Dorado, "Automated detection of load changes in large-scale networks," in *Proceedings of TMA*, 2009, pp. 34–41.
[4] E. Dagum and S. Giannerini, "A critical investigation on detrending procedures for non-linear processes," *J. Macroecon.*, vol. 28, no. 1, pp. 175–191, 2006.
[5] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area Internet traffic patterns and characteristics," *IEEE Netw.*, vol. 11, no. 6, pp. 10–23, Nov/Dec 1997.
[6] P. Heegaard, "Evolution of traffic patterns in telecommunication systems," in *Proceedings of CHINACOM*, 2007, pp. 28–32.
[7] F. Mata, J. García-Dorado, and J. Aracil, "Multivariate fairly normal traffic model for aggregate load in large-scale data networks," in *Proceedings of WWIC*, 2010, pp. 278–289.
[8] R. Birke, *et al.*, "Experiences of VoIP traffic monitoring in a commercial ISP," *Int. J. Netw. Manag.*, vol. 20, no. 5, pp. 339–359, 2010.
[9] L. Brown *et al.*, "Statistical analysis of a telephone call center: A queueing science perspective," *J. Am. Stat. Assoc.*, vol. 100, pp. 36–50, 2005.
[10] W. Chen, H. Hung, and Y. Lin, "Modeling VoIP call holding times for telecommunications," *IEEE Netw.*, vol. 21, no. 6, pp. 22–28, 2007.
[11] Q. He, C. Dovrolis, and M. Ammar, "On the predictability of large transfer TCP throughput," *Comput. Commun. Rev.*, vol. 35, no. 4, pp. 145–156, 2005.
[12] J. Taylor, "Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles," *Int. J. Forecasting*, vol. 26, no. 4, pp. 627–646, 2010.
[13] A. Finamore, M. Mellia, M. Meo, M. Munafo, and D. Rossi, "Experiences of internet traffic monitoring with Tstat," *IEEE Netw.*, vol. 25, no. 3, pp. 8 –14, 2011.
[14] J. Durbin, "Weak convergence of the sample distribution function when parameters are estimated," *Ann. Stat.*, pp. 279–290, 1973.
[15] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*. New York: John Wiley & Sons Inc., 1990.
[16] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*, ser. Springer Series in Statistics. Springer, 1991.