Load Characterization and Anomaly Detection for Voice Over IP Traffic

Michel Mandjes, Iraj Saniee, *Member, IEEE*, and Alexander L. Stolyar

***Abstract*—We consider the problem of traffic anomaly detection in IP networks. Traffic anomalies typically arise when there is focused overload or when a network element fails and it is desired to infer these purely from the measured traffic. We derive new general formulae for the variance of the cumulative traffic over a fixed time interval and show how the derived analytical expression simplifies for the case of voice over IP traffic, the focus of this paper. To detect load anomalies, we show it is sufficient to consider cumulative traffic over relatively long intervals such as 5 min. We also propose simple anomaly detection tests including detection of over/underload. This approach substantially extends the current practice in IP network management where only the first-order statistics and fixed thresholds are used to identify abnormal behavior. We conclude with the application of the scheme to field data from an operational network.**

***Index Terms*—Anomaly detection, heavy-tailed holding times, load characterization, network management, second-order statistic, traffic measurements, voice-over IP.**

I. Introduction

IP NETWORKS carrying voice traffic are beginning to emerge due to the cost efficiency of IP platforms and their extensibility to other applications and media. However, before this transition becomes widespread, key problems in network management and operation need to be addressed. In this paper we focus on load characterization, overload detection and more generally *load anomaly detection* in segments of IP networks that carry (almost) exclusively voice traffic, e.g., an egress port of an IP router or switch connected to the trunking voice gateway. As it turns out, many other segments of an IP infrastructure carrying large amounts of voice traffic may be dedicated to carry mostly voice traffic. This emerges from the architectures of many IP-based networks carrying voice, as shown in Fig. 1.

The need for detection of load anomalies arises, for example, when an atypical load change (increase or decrease) is experienced by a portion of the network. In *focused overload* a large number of callers try to reach the same destination phone number(s) and the network admits too many calls. More generally, overload or indeed *underload* occurs when a segment of the network fails and the traffic either overflows into a normal segment of the network or migrates away from it. In any one of these scenarios, it is desirable to *detect* the ongoing overload or underload as fast as possible, thus, helping the network operator to take remedial action or to invoke programmed response. This goal has, unfortunately, turned out to be elusive for the IP networks for a variety of reasons and typically the network operator hears about performance problems from customers before the (large amounts of) information available to it has been adequately analyzed. We show that at least in the voice over IP (VoIP) segments of an IP network, this need not be the case.

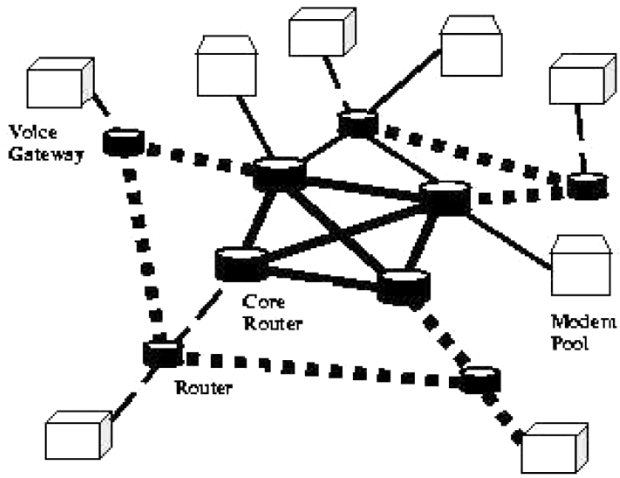

Fig. 1. Typical architecture of an IP network that carries a significant amount of VoIP traffic, with VoIP-only network segments highlighted in dashed line.

In an IP-based network, the traffic information available is the cumulative amounts of traffic ("byte counts") over 5-min time intervals [the *ifInOctets* and *ifOutOctets* management information base (MIB) in simple network management protocol (SNMP)]. For data traffic 5 min is indeed a long time. Can such "crude" information as 5-min byte counts be efficiently used for load anomaly detection? We show in this paper that for the VoIP traffic the answer is "yes," under a mild set of assumptions on the coding rate(s) of the packetized voice traffic and the mean call duration.

Our approach is based the analysis of the variance of the byte counts (i.e., second-order statistics.) Let $A(t_1, t_2)$ denote the cumulative amount of traffic sent on a link in the time interval (t_1, t_2). Expressions for the mean and variance of $A(t_1, t_2)$ can be derived for a very general model of IP traffic, with data sessions arriving as a Poisson process. (We do that in Section III-A). In the special case of a link with VoIP-only traffic, which is the focus of this paper, the mean and variance of $A(t_1, t_2)$ have very simple closed-form expressions, that are described in the Section III-B. As it turns out, simple tests

Manuscript received November 17, 2003; revised April 2, 2005.

M. Mandjes is with the Center for Mathematics and Computer Science (CWI), 1090 GB Amsterdam, The Netherlands (e-mail: michel@cwi.nl).

I. Saniee and A. L. Stolyar are with the Mathematical Sciences Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: iis@research.bell-labs.com; stolyar@research.bell-labs.com).

Digital Object Identifier 10.1109/TNN.2005.853427

based on these formulae allow us to discriminate between VoIP and non-VoIP traffic, and detect anomalies in the former.

There is a relatively large literature on anomaly detection in communication networks. For a recent summary, see [1]. For use of sample variance in conjunction with MIB variables, see [12]. For other approaches, e.g., Bayesian belief networks in conjunction with MIB variables, see [2]. Our approach differs from these approaches due to the derivation and validation of analytical formulae for the detection parameter (variance).

In summary, the contribution of the paper is the following:

- we argue that second-order statistics are useful in traffic anomaly detection and obtain expressions for the variance of a byte count for a quite general IP traffic model;
- we show that, in the case of VoIP traffic, byte counts over relatively long intervals can indeed be used for anomaly detection;
- we develop a set of detection procedures, for different types of anomalies;
- we assess the efficacy of the procedures with real traffic traces.

The paper is organized as follows. In Section II, we briefly discuss and contrast byte count traces for VoIP and general IP traffic from an operational network. This provides motivation for Section III in which we first derive a formula for the variance of byte count measurements for general IP traffic, and then present explicit formulae for the VoIP traffic with Pareto and exponential call duration distributions. In Section IV, we discuss usefulness of the measured sample byte count variance for detection of anomolous behavior. In Section V, we describe three principal types of alarms that can be generated with the collected traffic data, using both the theoretical and measured variances. In Section VI, we apply the technique presented to data collected from an operational network. Finally, in Section VII we provide a summary of the methodology proposed in this paper.

II. Measurements From an Operational Network

Before we describe a model for VoIP traffic, it is instructive to look at traces of VoIP and general IP traffic both measured in the same operational IP-based network. Figs. 2 and 3 show the 5-min byte counts on two distinct ports of a service provider network collected over one week. The first trace is taken from a port that carries VoIP traffic only, including IP signaling traffic and possibly other marginal non-VoIP load, and the second trace is taken from a port that carries general IP traffic such as WWW, TCP/IP, etc.

Both data sets were obtained via standard SNMP MIB agents with 5-min aggregation collected over a period of roughly one week. One outstanding feature of both data sets is the daily regularity of the load for both IP and VoIP as observed, for example, by Thomson *et al.* [11]. Also, simple visual inspection of the two profiles in Figs. 2 and 3 shows that the VoIP traffic is "smoother" and has less variability than the corresponding IP traffic trace. We show that this apparent regularity of the VoIP traffic can be more formally defined and exploited for detection of uncharacteristic (anomalous) load variation. In particular, we show that the observed variability in Fig. 2 is completely within the range

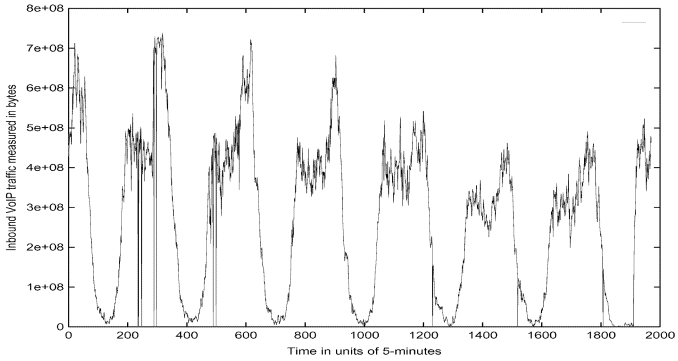

Fig. 2. Voice over IP traffic volume measured in 5-min intervals at an egress port of a T3 (45 Mb/s) trunk over a period of one week.

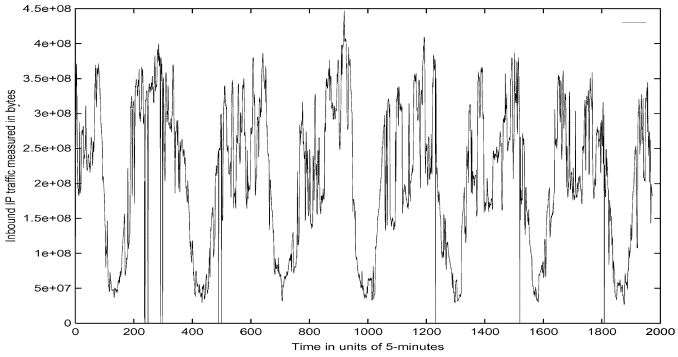

Fig. 3. Data-only IP traffic volume measured in 5-min intervals at an egress port of a T3 (45 Mb/s) trunk over a period of one week.

predicted by the proposed theoretical model, while that of Fig. 3 exceeds it.

III. Variance of an Interval Measurement

A. General Expression

Suppose *data sessions* (which in particular may be voice calls) arrive in time according to a (possibly nonhomogeneous) Poisson point process on the real axis $R = (-\infty, \infty)$. See [10] for the precise definition of a Poisson point process. Let Λ denote the *intensity measure* of this process, defined on Borel subsets of R. We assume that Λ is finite on bounded sets. This in particular means that the (random) number of arrivals (points) within any finite interval $(t_1, t_2]$ has Poisson distribution with mean $\Lambda\{(t_1, t_2]\}$. (When Λ is absolutely continuous with respect to Lebesgue measure, with the constant density $\lambda = d\Lambda/dt$, then this a homogeneous process with intensity λ.)

Assume that the (random) amount of traffic generated by a session arrived at time t is described by a random nondecreasing nonnegative right-continuous function

$$G_t(x), \quad x \geq 0$$

where $G_t(x)$ is the amount of traffic generated in the closed interval $[t, t + x]$. (Let $G_t(0-) = 0$ by convention.) We assume that the distribution of G_t depends on t as a parameter, and that the functions G_t corresponding to different sessions are independent (even if arrival times of some of them coincide).

Remark: Formally, the process we described is a *marked* Poisson point process [10], with the random functions G being

marks of the points. This means that the dependence of the distribution of G_t on t must be such that the marked point process can be well defined as a (measurable) random element. This condition is not restrictive for any conceivable application.

Let us denote for $0 \le x_1 \le x_2 < \infty$

$$f_t^{(1)}(x_1, x_2) \doteq \mathbf{E}[G_t(x_2) - G_t(x_1-)]$$
$$f_t^{(2)}(x_1, x_2) \doteq \mathbf{E}[G_t(x_2) - G_t(x_1-)]^2.$$

Now, let $A(t_1, t_2)$ denote the total amount of traffic generated by all sessions in the (closed) interval $[t_1, t_2]$.

Theorem: For the mean total amount traffic we have

$$\mathbf{E}A(t_1, t_2) = \int_{-\infty}^{t_2} f_t^{(1)}((t_1 - t)^+, t_2 - t)\Lambda(dt) \quad (1)$$

where $x^+ \doteq \max\{x, 0\}$.

If $\mathbf{E}A(t_1, t_2) < \infty$, then

$$\mathbf{V}A(t_1, t_2) = \int_{-\infty}^{t_2} f_t^{(2)}((t_1 - t)^+, t_2 - t)\Lambda(dt). \quad (2)$$

Proof Sketch: Suppose $X_i, i = 1, 2, \ldots$, is a sequence of independent identically distributed (i.i.d.) random variables, and N is random variable with Poisson distribution, independent of the X_i's. It is well known that

$$\mathbf{E}\left(\sum_{i=1}^{N} X_i\right) = \mathbf{E}N \cdot \mathbf{E}X_1 \quad (3)$$

and

$$\mathbf{V}\left(\sum_{i=1}^{N} X_i\right) = \mathbf{E}N \cdot \mathbf{E}X_1^2. \quad (4)$$

It follows directly from the definition of a marked Poisson point process, that in our case the total contributions into $A(t_1, t_2)$ of the sessions originating in nonoverlapping intervals $(a, b]$ and $(c, d]$ are independent. Therefore, it is sufficient to prove the (1) and (2) for the case when the intensity measure is concentrated on a finite interval $(a, b]$.

In this case, our marked point process can be constructed as follows. A Poisson random variable N with mean $\Lambda\{(a, b]\}$ is defined. Also, a sequence of i.i.d. random "extended" marks is defined, where each mark is a random function $((G_t(x), x \ge 0), t \in (a, b])$ of *two* variables x and t. A realization of our process is constructed by first taking a realization of N, then placing N points into the interval $(a, b]$ independently according to the distribution $\Lambda(dt)/\Lambda\{(a, b]\}$, and finally "attaching" first N extended marks (from the i.i.d. sequence) to the points. The mark $(G_t(x), x \ge 0)$ of a point located at t is simply the projection of its extended mark at time t.

Then, formulas (1) and (2) follow directly from (3) and (4), respectively. □

B. VoIP Formulas

Suppose now that the data sessions are voice calls with i.i.d. durations, having the distribution function $H(x), x \ge 0$, with the density $h(x), x \ge 0$, and finite mean μ^{-1}. Assume that calls arrive according to a homogeneous Poisson process with intensity $\lambda > 0$, and that each call in progress generates data at the constant rate 1.

In this case the distribution of the amount of traffic $A(t_1, t_2)$ depends only on $t_2 - t_1$, so we will write $A(t) \doteq A(0, t)$, and obviously $\mathbf{E}A(t) = (\lambda/\mu)t$. Although the general formula (2) can be applied directly to obtain $\mathbf{V}A(t)$, in our special case the derivation can be somewhat simplified by using the well known fact that (in a $M/GI/\infty$ queueing system) the random number N_1 of calls in progress at time 0 has Poisson distribution with mean λ/μ, and the residual call durations are i.i.d. with the density $h^*(x) = \mu(1 - H(x))$ and the distribution function $H^*(x) = \int_0^x h^*(y)\,dy$. Also, the number of calls which originate in $[0, t]$ is Poisson with mean $N_2 = \lambda t$.

If we denote $A_1(t) :=$ traffic generated in $[0, t]$ by flows that were already present at time 0, and $A_2(t) :=$ traffic generated in $[0, t]$ by flows that entered during the interval, then

$$\mathbf{V}A(t) = \mathbf{V}A_1(t) + \mathbf{V}A_2(t)$$

where [applying (4)]

$$\mathbf{V}A_1(t) = (\lambda/\mu)\int_0^t u^2 h^*(u)\,\mathrm{d}u + t^2(1 - H^*(t))$$

and

$$\begin{aligned}\frac{\mathbf{V}A_2(t)}{\lambda t} &= \int_0^t \int_0^{t-u} \frac{s^2}{t} h(s)\,\mathrm{d}s\,\mathrm{d}u \\ &\quad + \int_0^t \int_{t-u}^{\infty} \frac{(t-u)^2}{t} h(s)\,\mathrm{d}s\,\mathrm{d}u \\ &= \int_0^t \int_0^{t-u} \frac{s^2}{t} h(s)\,\mathrm{d}s\,\mathrm{d}u \\ &\quad + \int_0^t \frac{(t-u)^2}{t}(1 - H(t-u))\,\mathrm{d}u.\end{aligned}$$

Consider two special cases.

- In the first, the call duration is exponential, i.e., $h(x) = \mu\exp(-\mu x)$. Straightforward calculations yield

$$\mathbf{V}A(t) = \frac{2\lambda t}{\mu^2} - \frac{2\lambda}{\mu^3}(1 - e^{-\mu t}). \quad (5)$$

  If we denote by $\bar{A}(t) \doteq A(t)/t$ the (random) average number of calls in an interval of length t, by $a = \lambda/\mu$ the mean number of calls, and by $T = \mu t$ the measurement interval length normalized by the mean call duration, then we can rewrite (5) in a "normalized" form:

$$\mathbf{V}\bar{A}(t) = \sigma_T^2(a) \doteq 2a\frac{e^{-T} - 1 + T}{T^2}. \quad (6)$$

  This is the formula found by Riordan [9]. (The extension to the case when there is only a finite number of "trunks" available for calls was obtained by Beneš [4].)
- Now assume that a call duration has a Pareto distribution. More precisely, for $s \ge 0$

$$1 - H(s) = \frac{1}{(s+1)^{\alpha}}, \quad 1 - H^*(s) = \frac{1}{(s+1)^{\alpha-1}}.$$

  With similar calculations

$$\mathbf{V}A(t) = \frac{2\lambda}{(1-\alpha)(2-\alpha)(3-\alpha)}\left(1 - \frac{1}{(t+1)^{\alpha-3}}\right) + \frac{2\lambda t}{(1-\alpha)(2-\alpha)}.$$

Related calculations for Gaussian sources were done in [3].

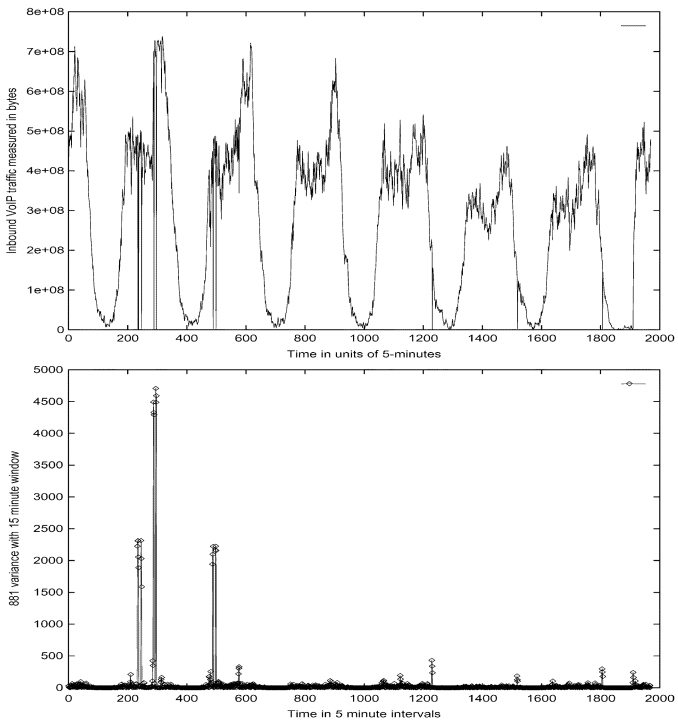

Fig. 4. VoIP traffic volume (top figure) and its normalized variance (measured in units of calls) calculated using a window size of six 5-min intervals (30 min).

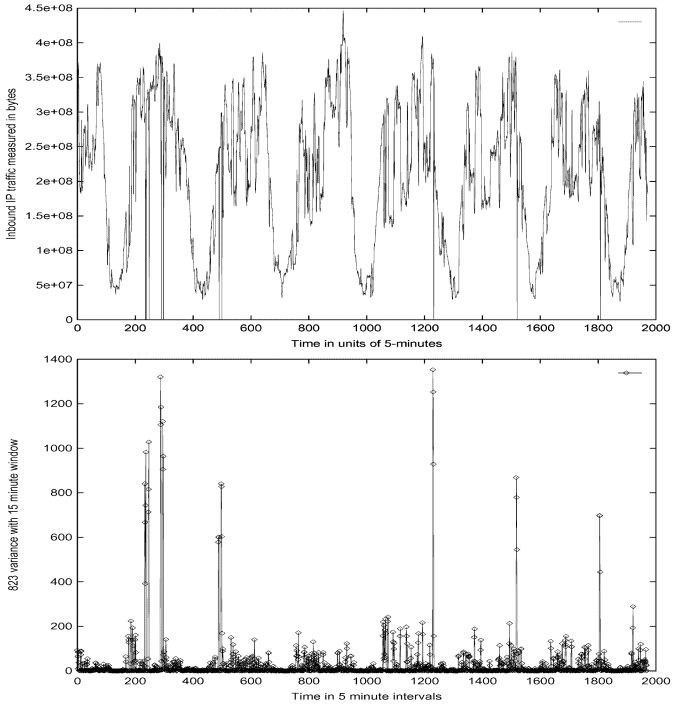

Fig. 5. Data-only IP traffic volume (top figure) and its normalized variance (measured in units of calls) calculated in a window size of six 5-min intervals (30 min) Profile of the IP traffic (top figure) and its normalized variances (to call count) calculated with a window of size 6 (30 min).

IV. Relevance of Variance in Setting Alarms

IP network management systems typically use no more than *static thresholds* to detect overload. While a first-order statistic, such as the average load, can provide useful information for non-peaked data, higher order statistics can provide much more refined information for even a richer variety of time series. For example, the peaks of the variance of counts in a fixed-sized moving window within a time series provides an indication of anomalous behavior in the underlying process, both for atypically high as well as low byte counts. To illustrate this with our data, we plot the variance (of the rescaled) byte counts with a 6-data-points window (30 min) of the VoIP traffic and compare it with the profile of the time series itself, that has some obvious anomalies, as shown in Fig. 4.

As it can be seen, the trivial anomalous behavior corresponding to call-count-drops-to-zero are (extremely) well detected by the peaks in the (moving) variance. The same holds true for the IP traffic (Fig. 5) although the spectrum here is fuller than that of VoIP.

The utility of second-order statistics for further detection of anomalies, both peaks and troughs, is therefore clear. However, second and higher order statistics provide a useful measure as long as their range of variability is known or at least can be predicted. This is what the variance formulas derived in the previous section and in particular Riordan's formula (6) provide.

V. Alarms for VoIP Load Anomalies

Consider a VoIP link, and assume call durations do have exponential distribution. Assume also that each call generates traffic at a constant rate γ, and that the mean call holding time $1/\mu$ is known. The length of a time interval over which byte counts are collected is t. (For all the examples in this paper, $t = 5$ min.) Then, Riordan formula provides an analytical expression for the variance of the byte counts, which is useful in detecting high load and abnormal traffic behavior. Based on this formula, we construct a set of traffic anomaly tests. (For more modern uses of the sample variance for detection of anomalies and abrupt changes in network traffic, see [6], [8], [12].)

Typically, three conditions need to be checked when one applies a model to (traffic) data. First, the applicability of the model to measurements needs to be tested. This is to check that the data meet the assumptions made in the model. Second, a test needs to determine if and when measurements indicate normal or abnormal load behavior, given that the data pass the first test. Third, a final test checks for over/underload. These tests, or *alarms*, are detailed in the following.

A. Traffic Model Alarm (Type I)

This alarm is issued when the empirical variances taken over a sliding window of the sequence of byte-counts $M_i, i = 1, 2, \ldots$ (where i indexes the consecutive 5-min intervals), differ substantially from the theoretical variance predicted by the Riordan formula. Such an alarm indicates *consistency* or "conformance" of the traffic statistics with those of VoIP traffic. Possible reasons for traffic "nonconformance" are the following.

- A significant fraction of the traffic is *not* VoIP. Evidently, if user data sessions do not generate a constant rate traffic (as voice calls do), Riordan formula does not apply.
- Even if traffic is VoIP, Riordan formula may predict a "wrong" variance, if the call holding time is not exponential or the actual mean call duration is different from what we assume it is.

- The test might fail due to the inherent nonstationarity of data. This problem could be circumvented by removing the trends in the data. This can be done in a straightforward way.
- Evidently, nonconsistency can be caused by continual misleading or false measurements.

If this alarm is indeed issued, the Type II and Type III alarms described in the subsections in the following, which *assume* measurements are VoIP, should be ignored because the current test indicates lack of agreement between observed measurements and the model.

To be precise, we propose the following procedure.

- Consider the sequence of byte-counts $M_i, i = 1, 2, \ldots$ (where, as before, i indexes the consecutive 5-min intervals), over a long observation interval I, typically many hours or days. This is the *training* or *learning* interval for characterization of data at hand.
- For each i, consider a *sliding window* consisting of n (typically 6–12) byte counts $M_{i-n+1}, \ldots, M_i$. Let $\bar{M}_i = (1/n)\sum_{k=i-n+1}^{i} M_k$ be their average. Compute the (normalized) empirical variance

$$\hat{\sigma_i^2} := \frac{1}{t^2}\frac{1}{\gamma^2}\frac{1}{n-1}\sum_{k=i-n+1}^{i}(M_k - \bar{M}_i)^2$$

  where γ is the data rate generated by one call in progress. The empirical variance $\hat{\sigma_i^2}$ should be close to the theoretical value $\sigma_T^2(a)$, given by the Riordan formula (6), where we set a to its estimate $\bar{M}_i/(\gamma t)$. (Note that it may be necessary to remove obvious linear trends, normalize the "noise" term in the regression model and generally clean the data within each sliding window before calculating the variance. This was done, for example, to obtain Figs. 6 and 7. These are standard statistical techniques which we will not elaborate here.)
- For each i, compare the empirical variance $\hat{\sigma_i^2}$ to $\theta\sigma_T^2(a)$, where θ is a fixed parameter, typically $1 \leq \theta \leq 2$. An event $\hat{\sigma_i^2} > \theta\sigma_T^2(a)$ we will call a *violation*. Obviously, frequent violations indicate that the VoIP traffic assumptions do not hold. Therefore, we issue the alarm if *the frequency of violations is too high* or *times intervals without violation are too short*. ("Too high" and "too short" is specified by additional parameters.)

Fig. 8 illustrates the procedure. Notice also that if n is chosen too small, the empirical variance will not be a reliable estimator of actual variance, whereas if n is chosen too large, the estimator may be bad due to nonstationarity of the data, i.e., because the individual observations are not likely to stem from the same distribution.

Figs. 6 and 7 show the result of the previous procedure for the VoIP and IP traffic, respectively, with the sliding window size of 6 data points (30 min) and $\theta = 1$. For the ease of comparison, the empirical variances are rescaled so that the corresponding theoretical variance is always $\sigma_T^2(1)$. To eliminate the natural trends present during various periods of the day, linear regression was used to "normalize" the data.

As it can be observed from these figures, the number of violations for the VoIP traffic is substantially smaller (5.3%) than

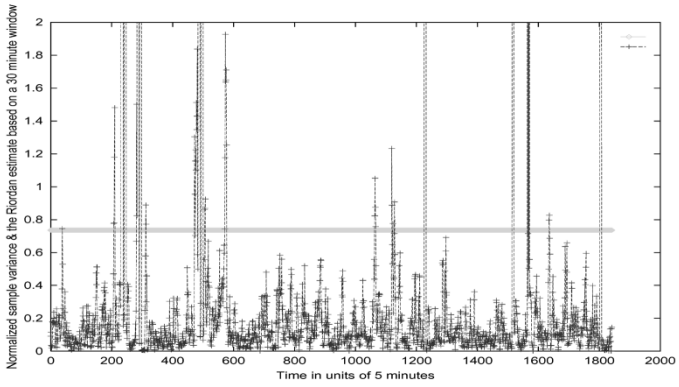

Fig. 6. Comparison of the sample variance with a window size 6 to the theoretical (Riordan) variance for the normalized VoIP traffic.

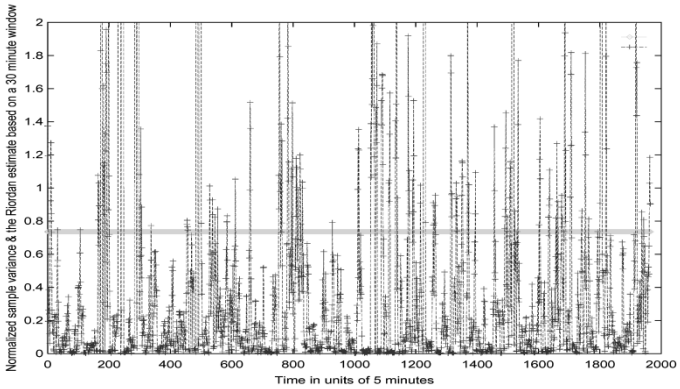

Fig. 7. Comparison of the sample variance with a window size 6 to the theoretical (Riordan) variance for the normalized IP traffic.

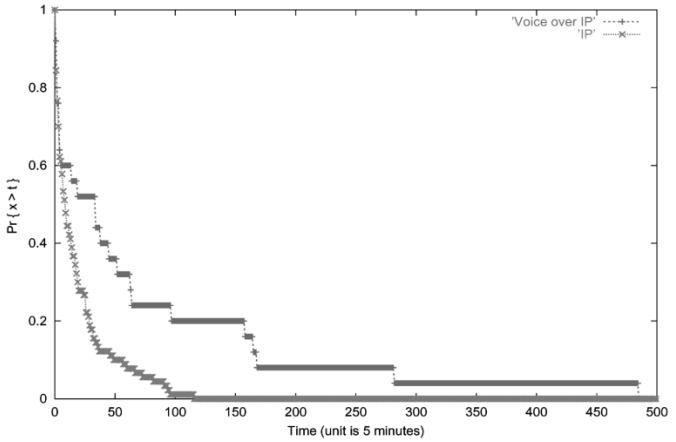

Fig. 8. Complementary cumulative distributions of the time intervals, in 5-min units, between consecutive violations for VoIP (top) and IP (bottom), see Figs. 6 and 7.

that of IP traffic (14.3%). This is in agreement with the fit of the theoretical model to VoIP and its lack of conformance to IP traffic (that is more bursty). This would result in setting of the Type I alarm for IP traffic and not for VoIP, as expected.

B. Fast Load Change Alarm (Type II)

This alarm is issued when the current byte count M_i *either above or below*, the (say) 95% confidence interval predicted by

Riordan formula with the mean load a set to the *short-term* empirical mean load taken over 2–4 last measurements. As noted before, this procedure has to be executed only if Type I alarm is not issued, in other words, there is no suspicion that the traffic is not VoIP.

This alarm may be caused by a number of events. A first possible cause is a link failure somewhere in the network. This failure triggers rerouting of calls, thus, leading to an effective load increase or decrease. Another reason may be a sudden traffic increase due to an "external" event, i.e., an event unrelated to the network. Also a measurement failure might lead to an alarm.

From extensive measurements of telephone traffic, we observe that the load in ∼15 min intervals can be treated as a stationary process. Taking this as an assumption, the procedure for issuing the alarm is as follows.

- Similarly to Type I test, compute averages $\bar{M}_i$ of byte counts M_i over a sliding window of size n. Typical value for n is from 2 to 4.
- Again as in Type I test, we assume that $\hat{a} \approx \bar{M}_i/(\gamma t)$, so the theoretical variance of a byte count is $\sigma_T^2(\hat{a})$.
- Construct a confidence interval $(\bar{M}_i - 2\sigma_T^2(\hat{a}), \bar{M}_i + 2\sigma_T^2(\hat{a}))$. If the new measurement M_{i+1} is outside this interval, issue the alarm.

C. Link Overload Alarm (Type III)

This alarm indicates that the link load is too high. It is issued when the current byte count (or the derived number of calls) exceeds a calculated threshold value, $M > M^*$, where M^* is a predefined threshold, and the first alarm is not set. The value of M^* is static, it is a fixed function of the link speed C and we show in the following how it is computed. Typically, this alarm requires some immediate action as link overload would result in packet losses and consequent voice quality deterioration. We note that static threshold alarms, such as this, are the most commonly used in network management. The main difference being that the threshold in this case is derived from the model and its estimated parameters and not empirically set. In other words, the model gives a threshold that can be used as a guide for setting an empirical overload threshold.

Suppose that the link speed C is large enough, say at least the speed of a T3 port, which corresponds to about 672 calls at 64 kbit/s. This guarantees that at the loads of the order of the link speed, the distributions of M_i and the instantaneous data rate are approximately normal. Suppose, we are given a constant $\beta > 0$ which is the maximum acceptable probability of the *instantaneous* data rate exceeding the link speed C. (Typically, β is 0.01 or 0.05.) Then the maximal acceptable link load a^* (i.e., the mean number of calls in progress) is computed from

$$a^* + b\sqrt{a^*} = C/\gamma.$$

where b is the $(1-\beta)$-quantile of the standard normal distribution. Then the threshold M^* can be chosen, for example, from

$$M^*/(\gamma t) + b\sigma_T(M^*/(\gamma t)) = a^*.$$

VI. Corroboration via Field Measurements

To test the methods proposed in the previous sections, we obtained data from an operational IP-based network that carries both IP and VoIP traffic. The general architecture of the network is similar to that shown Fig. 1 with segments that carry primarily VoIP traffic. The data contains 5-min input–output byte measurements per interface (*ifInOctets* and *ifOutOctets* SNMP MIBs) collected over a period of one week. The resulting byte counts from the two typical interfaces, one for each of IP and VoIP loads, were shown in Figs. 2 and 3.

A few observations are in order here. First both data sets show a fair amount of time-of-day dependence and therefore nonnegligible trends at various times during each day. To avoid the problems associated with nonstationary data, and the applicability of the model presented in Section III, we use 15 min—or three consecutive measurements—as the maximum time during which data will be assumed stationary and for larger windows we will determine trends and remove them for further analysis. Second, based on the information from the VoIP type of service provided in this network, we assume the coding of voice is at 64 kb/s without silence suppression or header compression, which together with real time protocol (RTP), user data protocol (UDP), and IP overheads results in average bandwidth per VoIP call ∼128 kb/s. Third, we assume the average holding time of a call is 2.5 min.[1] Exploiting this fact, the byte count time series shown in Fig. 3 can be converted to a "5-min average call count" processes. Third, we will treat the data as if it were collected online and apply the techniques as such, without assuming that the whole set is available upfront.

We apply the tests proposed in Section V to the available traces. First, for the applicability of the method developed in Section III to these data (Type I Test), 30-min sample variances are normalized by removing any trends and rescaled to match $a = 1$ Erlang load. Fig. 6 shows the plot of the normalized sample variances of the VoIP process versus the normalized variance from Riordan formula (6), that is, value of this formula for a unit load $a = 1$. As it may be seen from this graph, there are ∼25 violations within the week, i.e., the events when the variance exceeds the predefined threshold. Also notice that the average time between the violations is ∼400 min (80 time units) units. Further, these intervals are spread to the right with a longer tail. In contrast, the IP traffic has ∼90 violations, with the average time between them ∼90 min (18 time units), with intervals closely clustered around the mean. These data are shown in detailed in Fig. 8, from which we conclude that VoIP traffic fits the variance estimate from (6) much better than IP traffic does.

Having passed Test of Type I for VoIP traffic, we proceed to set the over/underload band using Type II Test for VoIP trace. Fig. 9 shows the profile of the data together with a confidence band for this alarm. The confidence band is valid under the assumption of 2.5-min call holding time with averaging interval of 15–30 min. We observe that with an expectation of ∼5% of data falling outside this adaptive band, ∼6% of the VoIP measured

[1]Similar computations for coding rates of 32 kb/s through 64 kb/s and average holding times up to 5 min show similar results. For example, with these values the corresponding plots to Figs. 6–10, not shown, establish exactly the same results described here.

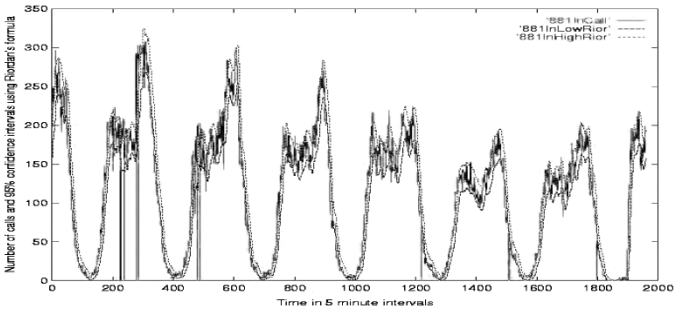

Fig. 9. 95% confidence band for VoIP traffic showing 6% of the data falls outside the band.

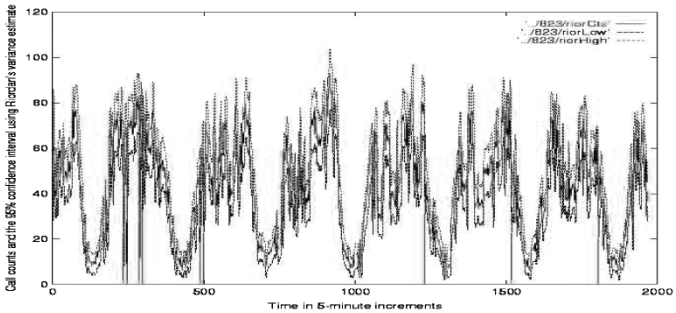

Fig. 10. 95% confidence band for the IP traffic, treating the IP traffic as if it were Voice over IP. A large fraction of the data points fall outside the confidence band.

data fall outside the window, again confirming the reasonableness of the fit of model to VoIP trace. (The confidence band for IP is shown in Fig. 10 for completeness.)

VII. Summary

We derived formulae for the variance of the cumulative traffic over fixed intervals for a very general model of data traffic. For voice over IP traffic, this formula is very simple and is known as the Riordan formula. It provides an (analytical) estimate for the variance of the VoIP load that passes through a switch or router interface. Standard router measurements (such as SNMP MIBS) also provide adequate data to estimate the variance of the traffic directly. These two estimates can then be used to determine if there is agreement between the model and data, and if so, provide an indication of load anomalies within the network segment where traffic is measured. The resulting method is also used to detect overload.

We examined the applicability of this scheme to a data set of field measurements of VoIP traffic and showed a good match between the analytical model and the measurements. To further test the usability of this scheme we also applied it to measurements from an IP interface that was shown to be much less consistent with the model. We classified the procedure into three tests or alarms. Type I alarm indicates that the traffic is unlikely to be VoIP. Type II alarm indicates anomalous load change and is applied only when alarm of Type I is not set. Finally, alarm of Type III is set only when there is overload. This alarm is also set only when alarm of Type I is not set.

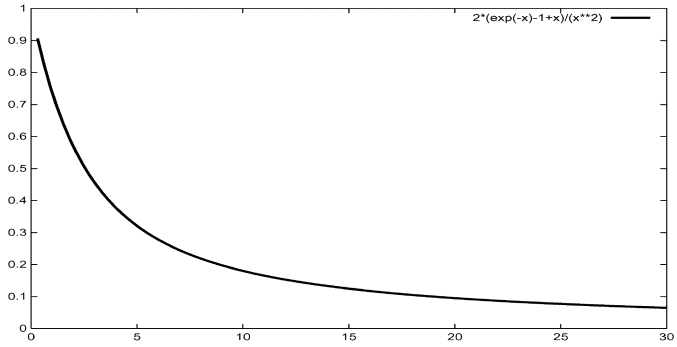

Fig. 11. Estimate of the standard deviation of the offered load of 1 erlangs as a function of the normalized aggregation interval $T = \mu t$, see formula (6). (For example, to read off the standard deviation for load of 500 erlangs, for the average holding time $1/\mu = 3$ min, and total load measured in bytes every $t = 5$ min, the value $\mathrm{T} = 5/3 = 1.666$ is read on the x axis, and the corresponding y axis value of $\sim$0.8 is multiplied by $500^{0.5} = 22.36$, giving the standard deviation of 18 calls.

We expect that the general variance formulae would be useful in detection of anomalies for the more general IP traffic. This is the subject of future research.

Appendix
Tradeoff Between Standard Deviation and Measurement Interval

Fig. 11 shows the plot of standard deviation of the offered load as a function of the normalized aggregation interval μT for a load of 1 erlangs in the Riordan formula. As it can be seen, the larger the aggregation interval, the better the estimate of the load,

$$\lim_{T \downarrow 0} \frac{e^{-T} - 1 + T}{T^2} = \frac{1}{2a}.$$

We see that with very short measurements the variance is just a (which is logical, as the number of calls X_t has a Poisson distribution with mean a and variance a). We also find that the variance decays to zero, essentially like $\sim T^{-1}$.

However, the need for a large interval for an accurate estimate of the variance needs to be balanced against the need to set alarms quickly when an anomaly is detected. For the latter, the shorter the aggregation interval, the better. The optimal tradeoff between these two tendencies depends on some quantification of the urgency of alarm sets versus accuracy of the alarms, given the Type I and Type II errors discussed earlier.

References

[1] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.
[2] C. Hood and C. Ji, "Proactive network fault detection," in *Proc. IEEE INFOCOM*, vol. 3, Kobe, Japan, Apr. 1997, pp. 1147–1155.
[3] R. Addie, P. Mannersalo, and I. Norros, "Performance formulae for queues with Gaussian input," in *Proc. ITC'16*, 1999, pp. 1169–1178.
[4] V. Benes, "The covariance function of a simple trunk group, with applications to traffic measurement," *Bell Syst. Tech. J.*, pp. 117–148, 1961.
[5] N. Duffield, "Queueing at large resources driven by long-tailed M/G/∞-modulated processes," *Queueing Syst.*, vol. 28, pp. 245–266, 1998.

[6] L. Ho, D. Cavuto, S. Papavassiliou, and A. Zawadzki, "Adaptive/automated detection of service anomalies in transaction-oriented WANS: Network analysis, algorithms, implementation, and deployment," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 5, pp. 744–757, May 2000.
[7] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
[8] R. Maxion and F. Feather, "A case study of ethernet anomalies in a distributed computing environment," *IEEE Trans. Reliab.*, vol. 39, no. 4, pp. 433–443, Oct. 1990.
[9] J. Riordan, "Telephone traffic time averages," *Bell Syst. Tech. J.*, vol. 39, pp. 1129–1144, 1951.
[10] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. New York: Wiley, 1995.
[11] K. Thompson, G. Miller, and R. Wilder, "Wide area internet traffic patterns and characteristics," *IEEE Network*, vol. 11, no. 6, pp. 10–23, Nov.-Dec. 1997.
[12] M. Thottan and C. Ji, "Proactive anomaly detection using distributed intelligent agents," *IEEE Network*, vol. 12, no. 5, pp. 21–27, Sep.-Oct. 1998.

Michel Mandjes received the M.Sc. degrees in both mathematics and econometrics and the Ph.D. degree from the Free University, Amsterdam, The Netherlands.

He worked as a Member of the Technical Staff at KPN Research and Bell Laboratories/Lucent Technologies, Murray Hill, NJ, and as a part-time Full Professor at the University of Twente. Currently, he has a joint position as the Department Head at the Center for Mathematics and Computer Science (CWI), Amsterdam, and as a Full Professor at the University of Amsterdam. His research interests include large deviations analysis of multiplexing systems, queueing theory, Gaussian traffic models, traffic management and control in IP networks, and pricing in multiservice networks.

Iraj Saniee (M'98) received the B.A., M.A., Tripos Part III, and Ph.D. degrees, all from Cambridge University, Cambridge, U.K.

He is currently the Director of the Mathematics of Networks and Systems Research Department at Bell Laboratories, Lucent Technologies, Murray Hill, NJ. The emphasis of research in his department is on the mathematical modeling, analysis, and optimization of emerging processes in wireless, data and optical networks. His recent work is on the control and optimization of resource-sharing networks, performance of limiting models of data traffic, and development of algorithms for the underlying mathematical and computational models. Prior to Bell Labs., he was at the Mathematical and Information Sciences Laboratories, Bellcore. He is currently an Associate Editor of *Operations Research*. He has published numerous articles in IEEE, INFORMS, and SIAM journals and proceedings.

Dr. Saniee is a member of IFIP WG 7.3.

Alexander L. Stolyar received the Ph.D. degree in mathematics from the Institute of Control Sciences, USSR Academy of Science, Moscow, Russia, in 1989.

He is currently a Member of Technical Staff in the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ. Before joining Bell Labs., he was with Institute of Control Sciences, Moscow, Russia; Motorola, Arlington Heights, IL; and AT&T Laboratories-Research, Murray Hill, NJ. His research interests are in stochastic processes, queueing theory, and stochastic modeling of communication, especially wireless, systems.