Describing Network Traffic Using the Index of Variability

Georgios Y. Lazarou
Department of Electrical & Computer Engineering
Mississippi State University
glaz@ece.msstate.edu

Julie Baca
Center for Advanced Vehicular Systems
Mississippi State University
baca@cavs.msstate.edu

Victor S. Frost, Joe Evans
Department of Electrical Engineering and Computer Science
University of Kansas
{frost, evans}@eecs.ku.edu

***Abstract*— Commonly used measures of traffic burstiness do not capture the fluctuation of traffic variability over the entire range of time scales. In this paper, we present a measure of variability, called the *Index of Variability* ($H_v(\tau)$), that fully and accurately captures the degree of variability (burstiness) of a typical network traffic process at each time scale and is analytically tractable for many traffic models. As an illustration, we derive the closed-form expression of $H_v(\tau)$ for two traditional traffic models and generate a variety of 2D and 3D Index of Variability curves. These curves demonstrate that the Index of Variability is a mathematically rigorous measure which can be used to fully characterize the complexities of the network traffic variability over all time scales. We then introduce a practical method for estimating the Index of Variability curve from a given traffic trace. Experimental results are presented which demonstrate the robustness of the method applied to the estimation of the Index of Variability curves from 12 NLANR network traffic long traces.**

I. Introduction

Many empirical studies have shown that Internet traffic exhibits high variability[1] [4][7][15][20]. That is, traffic is bursty (variable) over a wide range of time scales in sharp contrast to the assumption that traffic burstiness exists only at short time scales while traffic is smooth at large time scales [15]. High variability in traffic has been shown to have a significant impact on network performance [5][15]. The results from [5][9][13][17] show that knowledge of the traffic characteristics on multiple time scales helps to improve the efficiency of traffic control mechanisms. Importantly, the design and provision of quality-of-service-guarantees over the Internet requires the understanding of traffic characteristics, such as variability.

Since the publication of [15], the popular belief is that the high variability in traffic is due to the *long-range dependence*(LRD) property of the traffic processes. In general, a (weakly) stationary discrete-time real-valued stochastic process $Y = \{Y_n, n = 0, 1, 2, \ldots\}$ with mean $\mu = E[Y_n]$ and variance $\sigma^2 = E[(Y_n - \mu)^2] < \infty$ is long-range dependent if $\sum_{k=1}^{\infty} r(k) = \infty$, where $r(k)$ measures the correlation between samples of Y separated by k units of time. If $\sum_{k=1}^{\infty} r(k) < \infty$, then Y is said to exhibit *short-range dependence* (SRD).

Common traffic models with LRD are based on self-similar processes. In traffic modeling, the term self-similarity is usually used to refer to the *asymptotically second order self-similar* or *mono-fractal* processes [19]. The definition of asymptotically second order self-similarity is as follow [15]: assume that Y has an autocorrelation function of the form $r(k) \sim k^{-\beta} L(k)$ as $k \to \infty$, where $0 < \beta < 1$ and the function L is slowly varying at infinity, i.e., $\lim_{k\to\infty} \frac{L(kx)}{L(k)} = 1\ \forall x > 0$. For each $m = 1, 2, 3, \ldots$, let $Y^{(m)} = \{Y_n^{(m)}, n = 1, 2, 3, \ldots\}$ denote a

[1] Fluctuation of traffic as a function of time.

new aggregated time series obtained by averaging the original series Y over non-overlapping blocks of size m, replacing each block by its sample mean. That is, for each $m = 1,2,3,\ldots$, $Y^{(m)}$ is given by

$$Y_n^{(m)} = \frac{Y_{nm-m+1} + \cdots + Y_{nm}}{m} \qquad n \geq 1. \tag{1}$$

The new aggregated discrete-time stochastic process $Y^{(m)}$ is also (weakly) stationary with an autocorrelation function $r^{(m)}(k)$. Then, Y is called asymptotically second order self-similar with self-similar parameter $H = 1 - \frac{\beta}{2}$ if for all k large enough, $r^{(m)}(k) \to r(k)$ as $m \to \infty$. That is, Y is asymptotically second-order self-similar if the corresponding aggregated processes $Y^{(m)}$ become indistinguishable from Y at least with respect to their autocorrelation functions. By definition, asymptotically second order self-similarity implies LRD and vice versa [19].

The parameter H is called the *Hurst parameter*. For general self-similar processes, it measures the degree of “self-similarity”. For random processes suitable for modeling network traffic, the Hurst parameter is basically a measure of the speed of decay of the tail of the autocorrelation function. And if $0.5 < H < 1$, then the process is LRD, and if $0 < H \leq 0.5$, then it is SRD. Hence, H is widely used to capture the intensity of long-range dependence of a traffic process, the closer H is to 1 the more long-range dependent the traffic is, and vice versa [19].

There are several methods for estimating H from a traffic trace. One of the most widely used is the *Aggregated Variance* method: for successive values of m that are equidistant on a log scale, the sample variance of $Y^{(m)}$ is plotted versus m on a log-log plot [2][21]. By fitting a least-square line to the points of the plot and then calculating its slope, an estimate of the Hurst parameter is obtained as $\hat{H} = 1 - \frac{slope}{2}$.

Another very popular method is based on wavelets [33]. Given a traffic trace Y_n, the Hurst parameter can be estimated as follows. For each scale j, the wavelet energy $\mu_j = \frac{1}{N_j}\sum_{k=1}^{n_j} d^2(j,k)$ is plotted versus j on a semi-log plot (i.e., $log_2(\mu_j)$ vs. j). By fitting a least-square line to the points of the curve region that *looks* linear and then computing its slope α, H is estimated as $\hat{H} = \frac{\alpha+1}{2}$.

A. Need of a New Measure of Variability

Commonly used measures of traffic burstiness, such as the peak-to-mean ratio, the coefficient of variation of interarrival times, the indices of dispersion for intervals and counts, and the Hurst parameter, do not capture the fluctuation of variability over different time scales.

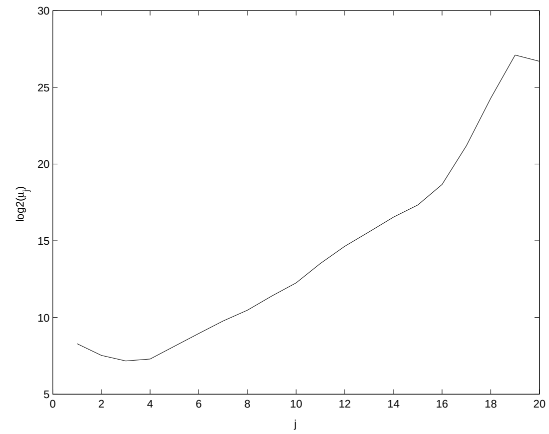

Fig. 1. $log_2(\mu_j)$ versus scale j for the Auckland-IV traffic trace 20010301-310-0 (For information about the trace, see Section IV).

It is claimed in [15] that the Hurst parameter is a good measure of variability, and the higher the value of H, the burstier the traffic. The popular belief from early studies [18][5][16][1] on the impact of LRD on network performance is that high values of the Hurst parameter are associated with poor queueing performance. But, later studies [12][13] show examples in which larger values of H are associated with better queueing performance compared to smaller values of H. In addition, the results in [17] indicate that the queueing performance depends mostly on the variability over certain time scales rather than on the value of H.

Moreover, it is known [9] that different long-range dependent processes with the same value of the Hurst parameter can generate vastly different queueing behavior. Clearly, the single value Hurst parameter does not capture the fluctuation of the degree of traffic burstiness across time scales, regardless if the traffic process exhibits LRD or SRD. From the definition presented in the previous section, the Hurst parameter is defined asymptotically (i.e., for large time scales) and hence conveys nothing about the variability of measured traffic over small or medium time scales; unless the traffic is exactly self-similar with known variances. Therefore, the Hurst paremeter is an incomplete descriptor of traffic variability.

For many network traffic processes, the wavelet energy-scale or variance-time plots usually do not tend to straight lines, i.e., see Figure 1. Usually many of these processes have piecewise fractal behavior with varying Hurst parameter over some small ranges of time scales [29]. Such processes are usually referred to as multifractal processes [34].

Queueing performace greatly depends on traffic irregularities at small time scales which are believed to be due to the complex dynamics of data networks [9][30]. Multifractal analysis based on the legendre spectrum is often used to study the multiscaling behavior of traffic at small time scales [29][35][36][37]. The process of estimating the legendre spectrum involves higher order sample moments and negative values of moments. It is known [31] that higher order sample moments are not well-behaved and negative values of moments tend to be erratic. In addition, the legendre spectrum is difficult to interpret [28].

Hence, there is a need for an intuitively appealing, coneptually simple, and mathematically rigorous measure which can capture the various scaling phenomena that are observed in data networks on both small and large scales [32].

In this paper, we present an alternative measure of variability, called the *Index of Variability* ($H_v(\tau)$), that fully and accurately captures the degree of variability of a typical network traffic process at each time scale and is analytically tractable for many traffic models.

The rest of this paper is organized as follows: In Section II, we present a rigorous definition of the *Index of Variability*. In Section III, as an illustration, we derive the closed-form expression of $H_v(\tau)$ for two traditional traffic models and generate a variety of 2D and 3D Index of Variability curves. In Section IV, we present a practical method for estimating the Index of Variability curve from a given traffic trace. We also present some experimental results to demonstrate the robustness of the method. We conclude the paper in Section V.

II. Index of Variability for Packet Traffic Sequences

Let $N(t)$ denote the number of events (packet arrivals) of a stationary point process in the interval $(0,t]$. For each fixed time interval $\tau > 0$, an event count sequence $Y = \{Y_n(\tau), \tau > 0, n = 1,2,\ldots\}$ can be constructed from each point process, where

$$Y_n(\tau) = N[n\tau] - N[(n-1)\tau] \tag{2}$$

denotes the number of events that have occurred during the n^{th} time interval of duration τ. Clearly, Y is also (weakly) stationary for all $\tau > 0$. In this study, Y represents a network traffic trace where $Y_n(\tau)$ denotes the number of packets observed from an arbitrary point in the network during the n^{th} time interval of duration τ. We refer τ as the *time scale* of the traffic trace, and it represents the length (i.e., 10ms, 1s, 10s, e.t.c.) of one sample of Y.

The expected number of events that have occurred during the interval $(0,t]$ is always: $E[N(t)] = \frac{t}{E[X]} = \lambda t$ where $E[X]$ is the expected interarrival time and λ is the mean event (packet) arrival rate. The index of dispersion for counts (IDC) is defined as: $IDC(t) \equiv \frac{Var[N(t)]}{E[N(t)]} = \frac{Var[N(t)]}{\lambda t}$. The IDC was defined such that it provides some comparison with the Poisson process, for which $IDC(t) = 1\ \forall t$. Note that since the point process is stationary, IDC has the same value over any interval of length t. Hence, t can be viewed as the time scale τ of the traffic process Y defined in (2). From now on we will be using t to denote generality and τ to denote time scales, i.e., the time length of each sample of the packet-count sequence Y.

An important feature of IDC is that it is mathematically equivalent to the Aggregated Variance method for estimating the Hurst parameter H of a self-similar process. For a self-similar process, plotting $log(IDC(m\tau))$ against $log(m)$ results in an asymptotic straight line with slope $2H-1$. When Y is a long-range dependent process, the slowly decaying variance property of LRD processes [15] with parameter $0 < \beta < 1$ is equivalent to an IDC curve[2] with an asymptotic straight line with slope $1-\beta$, implying $0 < slope < 1$. When the IDC curve converges to an asymptotic straight line with $slope = 0$ for some $\tau < \infty$, then Y is a short-range dependent process. Based on the above property of IDC, we define the following new measure of variability:

Definition 1: For a general stationary traffic process Y as defined by (2) whose $IDC(\tau)$ is continuous and differentiable over $(0,\infty)$, we call

$$H_v(\tau) \equiv \frac{\frac{d(log(IDC(\tau)))}{d(log(\tau))} + 1}{2} \tag{3}$$

the *index of variability* of Y for the time scale τ, where $\frac{d(log(IDC(\tau)))}{d(log(\tau))}$ is the local slope of the IDC curve at each τ when plotted in log-log coordinates.

Note that the index of variability is so defined in order that for a long-range dependent (asymptotically or second-order self-similar) process $H_v(\tau) = H \in (0.5,1)$ for all $\tau \geq \tau_o > 0$. The value of τ_o depends on the particular process. If the process is exactly self-similar then $H_v(\tau) = H \in (0.5,1)$ for all $\tau > 0$. That is, if $log(IDC(\tau))$ is linear with respect to $log(\tau)$), then $H_v(\tau)$ reduces to H. The Index of Variability can be thought of as the Hurst parameter defined at each time scale.

In general[3], the process Y exhibits significant variability for those time scales τ such that $0.5 < H_v(\tau) < 1$. When $\frac{d(log(IDC(\tau)))}{d(log(\tau))} \to 1$, then $H_v(\tau) \to 1$ implying very high variability. A plot of $H_v(\tau)$ versus τ would

[2] In log-log coordinates.

[3] The generality here is confined for those processes that are suitable in modeling network packet traffic.

depict the behavior of the traffic process Y in terms of variability (burstiness) at each time scale τ (= $10ms$, $100ms$, $1s$, $\ldots$).

Expanding the local slope of the IDC curve at each time scale, we get

$$\begin{aligned}\frac{d(log(IDC(\tau)))}{d(log(\tau))} &= \frac{\tau}{IDC(\tau)}\frac{d(IDC(\tau))}{d\tau} \\ &= \frac{\tau}{Var[N(\tau)]}\frac{d(Var[N(\tau)])}{d\tau} - 1. \quad (4)\end{aligned}$$

Using the above in (3), we obtain a more convenient form of the Index of Variability:

$$\begin{aligned}H_v(\tau) &= 0.5\tau\left(\frac{\frac{dVar[N(\tau)]}{d\tau}}{Var[N(\tau)]}\right) \quad (5)\\ &= \frac{1}{2}\left\{1+\tau\left(\frac{\frac{d(IDC(\tau))}{d\tau}}{IDC(\tau)}\right)\right\} \quad (6)\end{aligned}$$

In addition, setting $\tau = mT$, where $T > 0$ and $m = 1,2,\cdots$, and using the relation $Var[Y^{(m)}] = \frac{Var[N(mT)]}{m^2}$, we can express the index of variability function in terms of $Var[Y^{(m)}]$ versus m:

$$H_v(mT) = 0.5m\frac{\frac{dVar[Y^{(m)}]}{dm}}{Var[Y^{(m)}]} + 1. \quad (7)$$

Suppose now Y is an aggregate sequence of packet counts resulting from the superposition of M independent packet-traffic sources, not necessarily identical. Then $N(t) = N_1(t) + \cdots + N_M(t)$, where $N_i(t)$ denotes the number of packet arrivals in the interval $(0,t]$ from the i^{th} traffic source. Assuming again stationarity, we have

$$IDC(t) = \frac{\sum_{i=1}^{M} Var[N_i(t)]}{\sum_{i=1}^{M}\lambda_i t} = \sum_{i=1}^{M}\left(\frac{IDC_i(t)}{\Lambda_i}\right) \quad (8)$$

where λ_i is the mean packet arrival rate from the i^{th} source, and $\Lambda_i = \frac{\sum_{j=1}^{M}\lambda_j}{\lambda_i}$. In addition, $\frac{log(IDC(t))}{log(t)} = \frac{log(\sum_{i=1}^{M} Var[N_i(t)])}{log(t)} - \frac{log(\sum_{i=1}^{M}\lambda_i t)}{log(t)}$, and upon taking the derivative in respect to log(t) we get the index of variability for the aggregate traffic stream to be

$$\begin{aligned}H_v(\tau) &= 0.5\tau\left(\frac{\sum_{i=1}^{M}\frac{dVar[N_i(\tau)]}{d\tau}}{\sum_{i=1}^{M} Var[N_i(\tau)]}\right)\\ &= \frac{1}{2}\left\{1+\tau\left(\frac{\sum_{i=1}^{M}\frac{d(IDC_i(\tau))}{d\tau}\left(\frac{1}{\Lambda_i}\right)}{\sum_{i=1}^{M}\left(\frac{IDC_i(\tau)}{\Lambda_i}\right)}\right)\right\}. \quad (9)\end{aligned}$$

As we can observe from (9), the variances or the indices of dispersion for counts of the M independent point-processes completely characterize the variability function of the aggregate packet-count sequence Y. If $\lim_{\tau\to\infty} IDC(\tau) = \lim_{\tau\to\infty}\left(\sum_{i=1}^{M}\left(\frac{IDC_i(\tau)}{\Lambda_i}\right)\right) = c < \infty$, then obviously, $\lim_{\tau\to\infty} H_v(\tau) = 0.5$. In case that all M underlying point processes of making up Y are also identical, then (9) reduces to (6). If all M underlying point processes are Poisson, then $\frac{d(IDC_i(\tau))}{d\tau} = 0$ for all τ and i and hence $H_v(\tau) = 0.5$ for all τ.

III. Analysis of Traffic Models in Terms of the Index of Variability

In this section, we derive the Index of Variability functions for two traditional traffic models: two-state Markov Modulated Poisson Process (MMPP) and renewal process with hyperexponential interarrival time distributions of order two (RPH2). Two-state MMPP models became popular for modeling the superposition of packet voice streams [11].

The work in [38] shows that long-tail distributions can be approximated by hyperexponentional distributions. Thus, renewal processes with hyperexponential interarrival time distributions can be used for capturing the high variability of traffic over any range of (short or long) time scales. A major advantage of these models is their relative ease of analytically obtaining queueing performance predictions.

A. Two-state MMPP

Here we consider that the underlying point process of Y is an MMPP with two-state Markov chain where the mean sojourn times in state 1 and 2 are α^{-1} and β^{-1}, respectively. When the chain is in state i ($i = 1,2$) the point process is Poisson with rate λ_i. Letting $\rho = \alpha+\beta$ and $\upsilon = \lambda_1\beta+\lambda_2\alpha$, we have from [11] that $E[N(t)] = \frac{\upsilon t}{\rho}$ and $IDC(t) = 1+\rho A - A\left(\frac{1-e^{-\rho t}}{t}\right)$, where $A = \frac{2\alpha\beta(\lambda_1-\lambda_2)^2}{\rho^3\upsilon}$. It is easy to see that $\lim_{t\to\infty} IDC(t) = 1+\rho A$. Upon taking the derivative of IDC(t) we obtain the index of variability of Y as

$$H_v(\tau) = 0.5\left\{1+\frac{A\left[1-(1+\rho\tau)e^{-\rho\tau}\right]}{(1+\rho A)\tau - A(1-e^{-\rho\tau})}\right\}.$$

1) Numerical Example: Assume $\alpha^{-1} = \beta^{-1} = 100$ seconds, $\lambda_1 = 4$ packets/second and λ_2 to vary from 1 to 1000 packets/second. Figure 2 shows the resulting index of variability curves as a function of time scale (τ) and state rates (λ_i). Notice that when $\lambda_2 = \lambda_1$, we have a pure Poisson process and therefore zero variability. But as the difference between λ_1 and λ_2 increases, so does the index of variability. From Fig. 2 we can observe that the index of variability increases with λ_2 up to its maximum value, and any further increase in λ_2 does not have any affect on variability. We can also observe that the index of variability increases with τ up to its maximum value and then decays exponentially. We can also notice that for values of λ_2 not very close to λ_1, the packet-count process Y has substantial variability over a wide range of time scales that spans about 200 seconds.

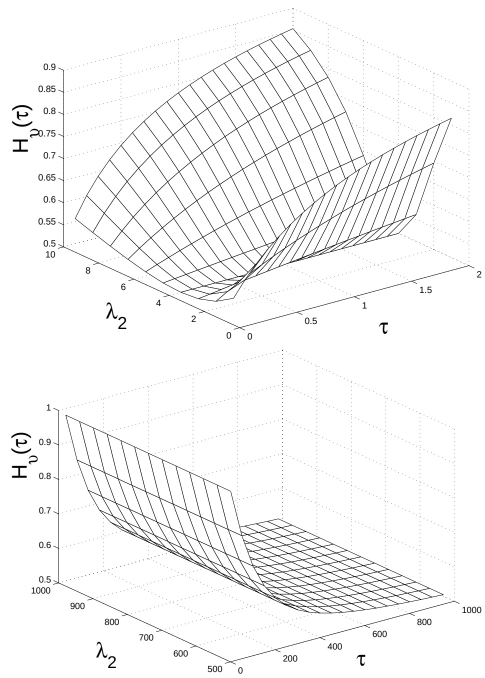

Fig. 2. Index of Variability for The Two-State MMPP: $\alpha^{-1} = \beta^{-1} = 100$ Seconds, $\lambda_1 = 4$ Packets/Second.

B. RPH2

We assume here that the underlying point process of Y is a stationary renewal process with interarrival times hyperexponentially distributed. We call this model as the *hyperexponential model*. A hyperexponential distribution of order K, $(= 1, 2, 3, \ldots)$, is the weighted sum of K exponential distributions:

$$F_K(x) = Pr[X \leq x] = \sum_{i=1}^{K} w_i (1 - e^{-\alpha_i x}) \qquad (10)$$

where $w_i > 0$ are the weights satisfying $\sum_{i=1}^{K} w_i = 1$, and $\alpha_i > 0$ are the rates of the exponential distributions [10]. It is shown in [8] that if $w_i = w^i$ and $\alpha_i = \frac{\mu}{\eta^i}$ for $0 < w < 1$, $\eta > 1$, and $\mu > 0$, then the tail of the hyperexponential distribution gets longer and longer with K. The major advantages of the hyperexponential distributions over heavy-tailed distributions like Pareto are two-fold: their Laplace transform exists, therefore they can be utilized in analytic models, and they have finite variance for all K.

In this paper, we only consider the case of $K = 2$. Letting $a = \alpha_1$ and $b = \alpha_2$, we get the pdf of the interarrival times to be:

$$f_2(x) = w_1 a e^{-ax} + w_2 b e^{-bx}. \qquad (11)$$

The mean packet arrival rate is $\lambda = \frac{ab}{aw_2 + bw_1}$, and the squared coefficient of variation of the interarrival times is $C^2(X) = 2\left[\frac{a^2 w_2 + b^2 w_1}{(aw_2 + bw_1)^2}\right] - 1$. Note that if $a = b$, then $\lambda = a = b$ and $C^2(X) = 1$ for all the values of w_1 and w_2, and hence we have a Poisson process. In addition, $\lim_{w_2 \to 0} C^2(X) = 1$ and $\lim_{b \to 0} C^2(X) = \frac{2}{w_2} - 1$. As shown in Fig. 3, for constant values of a and b, $C^2(X)$

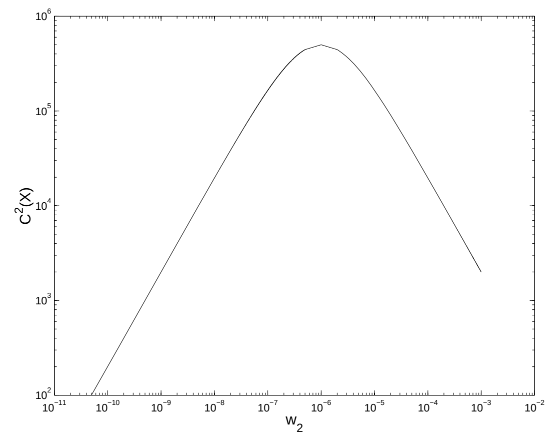

Fig. 3. Squared Coefficient of Variation of the Interarrival Time vs. w_2 for the Case of Hyperexponential Distribution of Order Two: $a = 100$, and $b = 0.0001$.

increases exponentially up to its maximum value and then decreases to one very abruptly. The maximum value depends on the value of b and it can get extremely high. This indicates that the hyperexponential distribution can be used to model the interarrival times distribution of highly bursty traffic.

From [3] we have that

$$Var[N(t)] = 2\lambda \int_0^t \Phi(u)\, du + \lambda t - \lambda^2 t^2 \qquad (12)$$

where

$$\Phi(t) = L^{-1}[\Phi^*(s)] = L^{-1}\left[\frac{f_2^*(s)}{s\{1 - f_2^*(s)\}}\right]. \qquad (13)$$

Note that the symbol L^{-1} denotes the inverse Laplace transform and

$$f_2^*(s) = L\left[f_2(x)\right] = w_1 \left(\frac{a}{s+a}\right) + w_2 \left(\frac{b}{s+b}\right)$$

is the Laplace transform of $f_2(x)$. Noting that

$$\begin{aligned} \varphi(t) &= L^{-1}\left[\frac{f_2^*(s)}{1 - f_2^*(s)}\right] \\ &= \lambda - \frac{[(aw_1 + bw_2)^2 - (a^2 w_1 + b^2 w_2)] e^{-[aw_2 + bw_1]t}}{aw_2 + bw_1} \end{aligned} \qquad (14)$$

we easily obtain $\Phi(t)$ as

$$\Phi(t) = \int_0^t \varphi(u)\,du = \lambda t - \frac{[(aw_1+bw_2)^2-(a^2w_1+b^2w_2)](1-e^{-[aw_2+bw_1]t})}{(aw_2+bw_1)^2} \qquad (15)$$

By performing the integration in (12) we get

$$Var[N(t)] = \frac{2\lambda[(aw_1+bw_2)^2-(a^2w_1+b^2w_2)]}{(aw_2+bw_1)^3}\left(1-e^{-[aw_2+bw_1]t}\right)+\lambda C^2(X)t, \qquad (16)$$

and hence

$$\frac{d}{dt}(Var[N(t)]) = \frac{2\lambda[(aw_1+bw_2)^2-(a^2w_1+b^2w_2)]}{(aw_2+bw_1)^2}e^{-[aw_2+bw_1]t}+\lambda C^2(X), \qquad (17)$$

and

$$IDC(t) = \frac{2[(aw_1+bw_2)^2-(a^2w_1+b^2w_2)]}{(aw_2+bw_1)^3}\left(\frac{1-e^{-[aw_2+bw_1]t}}{t}\right)+C^2(X). \qquad (18)$$

Observe that $\lim_{t\to\infty} IDC(t) = C^2(X)$, and if $a=b$ then $[(aw_1+bw_2)^2-(a^2w_1+b^2w_2)]=0$ and $C^2(X)=1$ making $Var[N(t)]=\lambda t$ and $IDC(t)=1$, i.e., we get a Poisson process. (16) or (18) can then be used in (5) or (6) to obtain the index of variability. It is obvious to see that $\lim_{\tau\to\infty} H_v(\tau)=0.5$.

Deriving the symbolic expression of $Var[N(t)]$ for $K>2$ is a difficult problem, mainly due to the difficulty in deriving $\phi(t)$, i.e., performing the following inverse Laplace transform:

$$L^{-1}\left[\frac{f_K^*(s)}{1-f_K^*(s)}\right]$$

where $f_K^*(s)$ is the Laplace transform of the the K-order hyperexponential pdf of the interarrival times. However, it becomes trivial when the model parameters (e.g., w_i and α_i) are set to numerical values.

1) Numerical Example: Let $a=100$. Table I lists the values of the mean packet rate (λ) and the squared coefficient of variation of the interarrival times ($C^2(X)$) for $b=0.01$ and $b=0.0001$ for different values of w_2. Note that $w_1+w_2=1$. Interesting, the maximum value of $C^2(X)$ occurs when $\lambda=\frac{a}{2}$. Also, Fig. 4 indicates that at this value of λ the process attains the widest range of time scales of high variability, and in this range the index of variability reaches its maximum value (curve (i), maximum$H_v=0.9988$). Observe that this widest range of time scales of high variability most likely covers all time scales that impact network performance evaluation [9]. In this example and for $\lambda=\frac{a}{2}$ packets/s, the range of time scales that the packet-count sequence Y exhibits high variability spans 7 order of magnitude.

TABLE I
VALUES OF MEAN PACKET RATE (λ) AND SQUARED COEFFICIENT OF VARIATION OF INTERARRIVAL TIMES ($C^2(X)$) FOR THE NUMERICAL EXAMPLE OF THE CASE OF HYPEREXPONENTIAL DISTRIBUTION OF ORDER TWO: $a=100$.

	λ	(packets/sec)	$C^2(X)$	
w_2	$b=0.01$	$b=0.0001$	$b=0.01$	$b=0.0001$
10^{-3}	9.1000	0.0999	1.6522x10^3	1.9950x10^3
10^{-4}	50.0000	0.9901	5.0000x10^3	1.9605x10^4
10^{-5}	90.9000	9.0909	1.6536x10^3	1.6529x10^5
10^{-6}	99.0000	50.0000	197.0202	5.0000x10^5
10^{-7}	99.9000	90.9091	20.9561	1.6529x10^5
10^{-8}	99.9900	99.0099	2.9992	1.9607x10^4
10^{-9}	99.9990	99.9001	1.2000	1.9970x10^3
10^{-10}	99.9999	99.9900	1.0200	200.9596
10^{-11}	100.0000	99.9990	1.0020	20.9996
10^{-12}	100.0000	99.9999	1.0002	2.9999
10^{-13}	100.0000	100.0000	1.0000	1.2001

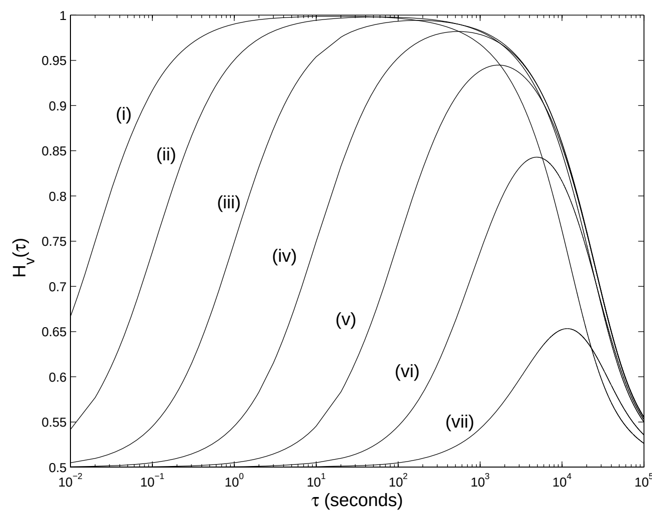

Fig. 4. Index of Variability vs. Time Scale for the Case of Hyperexponential Distribution of Order Two: $a=100$, $b=0.0001$, w_2 = (i) 10^{-6} (ii) 10^{-7} (iii) 10^{-8} (iv) 10^{-9} (v) 10^{-10} (vi) 10^{-11} (vii) 10^{-12}.

In addition, Figure 4 shows that the maximum value of variability as well as the range of time scales of substantial variability become smaller as $\lambda \to a$. Let

$$\tau_{on} = inf\{range\ of\ time\ scales\ of\ substantial\ variability\},$$

and

$$\tau_{off} = sup\{range\ of\ time\ scales\ of\ substantial\ variability\}.$$

As we can see from these curves, τ_{on} gets bigger as λ approaches a. Although it is not completely shown in Figure 4, it is not difficult to see that τ_{off} becomes smaller as $\lambda \to b$. Notice that for all $\tau \ni [\tau_{on}, \tau_{off}]$ the process looks like Poisson.

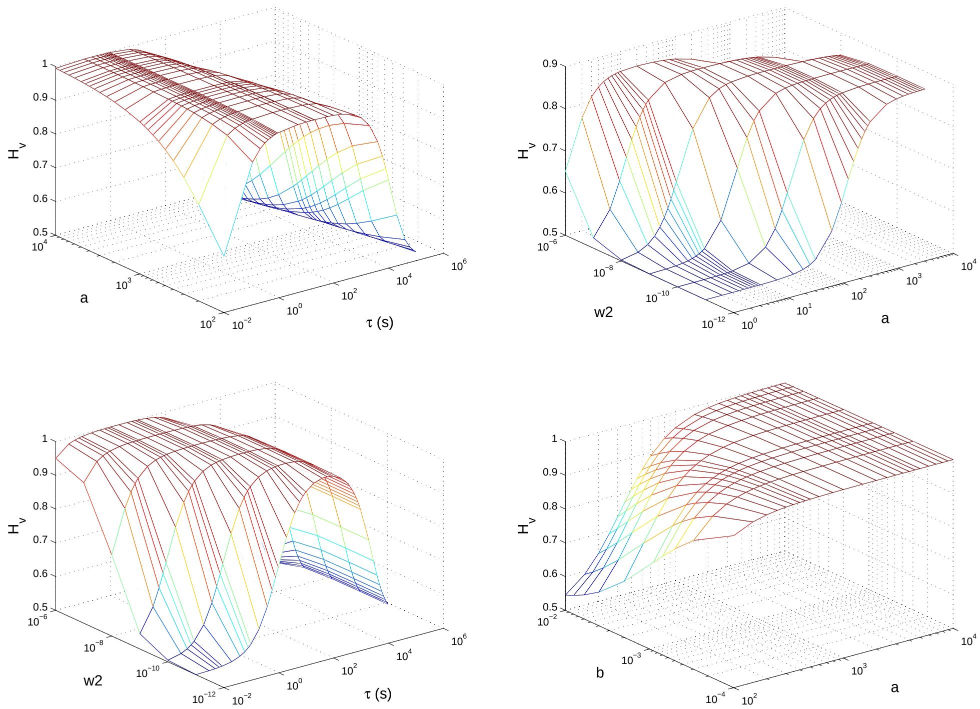

Fig. 5. Illustrations of 3D Index of Variability curves generated using the *hyperexponential* model with the following model parameter values: (top left) $b = 0.0001$, $w_2 = 10^{-6}$ (top right) $\tau = 1000$, $b = 0.001$ (bottom left) $a = 1000$, $b = 0.0001$ (bottom right) $\tau = 1$, $w_2 = 10^{-7}$.

Figure 5 depicts 3D Index of Variability curves generated using the *hyperexponential* model with $K = 2$. Clearly, both Figures 4 and 5 demonstrate that the *hyperexponential* model can yield a variety of Index of Variability curves. Hence, *hyperexponential* models can be used to model a wide range of network traffic types. Although *hyperexponential* models of order two (i.e., $K = 2$) are capable of generating a variety of Index of Variability curves, to capture the characteristics of traffic with multimodal Index of Variability curves it would be necessary to use higher order ($K > 2$) *hyperexponential* models.

IV. Estimating $H_v(\tau)$ from Traffic Traces

The estimation of the Index of Variability curve from a given traffic trace requires the estimation of the first derivative of $Var[N(\tau)]$ from discrete samples ($Var[N(\tau_i)]$, $i = 1, \cdots, n$). To do this, we must first find an analytic function that best fits the discrete variance data. This in turn requires the use of an interpolation method such as polynomial-based interpolation, cubic spline and smoothing spline [23]-[27].

Since we use the sample variances as the estimates of $Var[N(\tau_i)]$, $i = 1, \cdots, n$, we consider these estimates of the variances to be noisy samples. The smoothing spline interpolation methods are known to have optimal properties for estimating continuous functions and their derivatives from a finite number of noisy samples [24], [26], [27]. Note that nonsmoothing interpolation methods such as cubic spline have the characteristic that the estimated curve passes through all the given points. Hence, in case of noisy data, nonsmoothing interpolation methods yield rough curves and therefore erroneously high first derivatives.

A. Smoothing Spline Interpolation Method

For a given data series (x_i, y_i), $i = 1, 2, \cdots, n$, the smooth function $f(x)$ is the solution of the minimization problem

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \xi \int_{x_1}^{x_n} (f^{(k)})^2 \, du, \tag{19}$$

where ξ is the smoothing parameter and $f^{(k)}$ is the k^{th} derivative of f. If $k = 2$, then f is a cubic smoothing spline.

The first term in (19) is the residual sum of squares, an indicator of the goodness-of-fit of the spline curve to the data. In other words, it measures the degree of fidelity of the smoothing spline function to the data. The second term measures the roughness of the resulting smoothing spline curve. The roughness of a function can be characterized by its curvature. For example, if a function is a straight line, then its second derivative (and therefore, roughness) is zero. That is, the second term is a penalty term measuring how close the function is to a straight line.

The smoothing parameter ξ plays an important role. It weights two aspects: smoothness and fit. Large values of ξ give a smoother curve, while small values of ξ result in a closer fit.

B. Steps for Estimating $H_v(\tau)$ from Traffic Traces

We now present a practical method for estimating the Index of Variability from traffic traces. Assuming that a given traffic trace is a realization of a second-order ergodic point process whose variance curve is continuous and differentiable. We can estimate $H_v(\tau)$ of the process as follows.

- Using the *Aggregated Variance* method [2] estimate the variance-time sequence: $\widehat{Var}[N(\tau_i)],\ i = 1, \cdots, n$.
- Using an appropriate smoothing spline implementation estimate the smoothing spline $\widetilde{Var}[N(\tau)]$ from $\widehat{Var}[N(\tau_i)],\ i = 1, \cdots, n$.
- Using (5) estimate the Index of Variability $\widehat{H_v}(\tau)$.

We validated the accuracy of this process by estimating and matching the Index of Variability curves shown in Figure 4 from synthetically generated data using the *hyperexponential* traffic model.

C. Experimental Results

In this section we present some experimental results to demonstrate the robustness of the method described in the previous section.

Using the steps outlined in the previous section, we estimated the Index of Variability curve ($H_v(\tau)$) from 12 NLANR network traffic long traces [22]. The dates at which each trace was collected and their durations are listed in Table II. For more information about these traffic traces, see [22].

We used Matlab's spline toolbox to estimate all the smoothing splines. Its smoothing spline implementation is based on Reinsch's approach [24], [25]. Based on the input data, the algorithm computes the optimal smoothing parameter ξ such that the penalized residual sum of squares is less than a tolerance value $\varepsilon > 0$. In all cases we used the default value of k $(= 2)$ and $\varepsilon = 0.0001$.

Fig. 6 and 7 show the estimated index of variability curves from the 12 long packet traces. As expected, all traces have high variability. Interesting, all curves exhibit a transient rising behavior. A very important observation is that the index of variability curve is quite different from the curves of the Auckland traces. We also observe that these empiracally obtained curves are similar to the curves analytically obtained by the *hyperexponential* model in Section III-B (see Fig. 4).

V. Discussion

VI. Conclusion

All commonly used measures of traffic burstiness do not capture the fluctuation of variability over different time scales. Therefore, we developed a novel measure of variability, called the *index of variability* ($H_v(\tau)$), that fully and accurately captures the degree of variability of a typical network traffic process at each time scale and is analytically tractable for many traffic models.

We then discussed of how to estimated $H_v(\tau)$ from empirically measured network traffic traces. In this paper, we estimated the index of variability from 12 NLANR network traffic long traces. The results show that the traffic variability can exhibit a non-monotonic behavior. In addition, the results show that the index of variability can fully capture the multifractal behavior of traffic processes, especially at small time scales. The results also suggest that renewal processes with interarrival times hyperexponentially distributed are suitable for modeling such network traffic processes. We are currently working in developing a method of fitting analytically obtained index of variability curves from the hyperexponential model to the curves estimated from traffic traces.

References

[1] A. Baiocchi, N. Blefari Melazzi, M. Listanti, A. Roveri, R. Winkler, "Modeling Isssues on an ATM Multiplexer Within a Bursty Traffic Environment," INFOCOM'91, Vol.1.

[2] J. Beran, *Statistics for Long-Memory Processes. Monographs on Statistics and Applied Probability*, Chapman and Hall, New York, NY, 1994.

[3] D. R. Cox, *Renewal Theory*, London: Methuen, New York: J. Wiley & Sons, 1962.

[4] D. Duffy, A. McIntosh, M. Rosenstein, D. Wilson, "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks," IEEE *JSAC*, 12(3):544-551, 1994.

[5] A. Erramilli, O. Narayan, W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic," IEEE/ACM Trans. on Net., Vol.4, No.2, April 1996.

[6] P. M. Fiorini, *Modeling Telecommunication Systems with Self-Similar Data Traffic*, Ph.D. Dissertation, University of Connecticut, Sept. 9, 1998.

[7] H. J. Fowler, W. E. Leland, "Local Area Network Traffic Characteristics, with Implication for Network Congestion Management," IEEE JSAC, Vol. 9, No. 7, Sept 1991.

[8] M. Greiner, M. Jobmann, and L. Lipsky, "The Importance of Power-Tail Distributions for Telecommunications Traffic Modeling," Technical Report, Institute für Informatik, Technische Universität München, München, Germany, 1995.

[9] M. Grossgluser and J.-C. Bolot, "On the Relevance of Long-Range Dependence in Network Traffic," IEEE/ACM Trans. on Net., Vol.7. No.5, Oct. 1999.

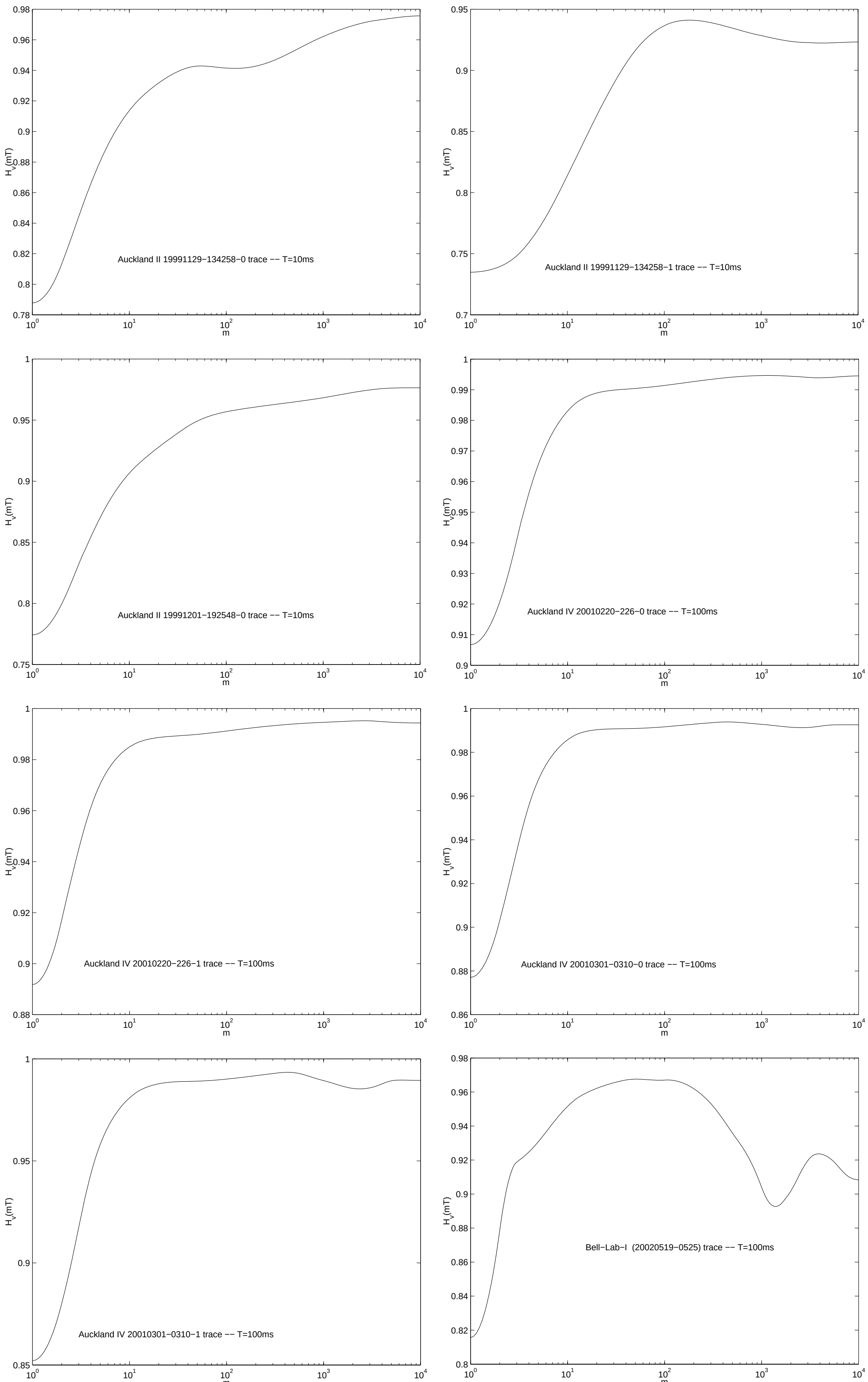

Fig. 6. Estimated Index of Variability Curves for Auckland II (top first three), Aucland IV (next four), and Bell-Lab-I (last) traces

TABLE II

NLANR NETWORK TRAFFIC TRACES

Trace	Data Set	Date Collected	Duration (Days:Hours:Minutes)
19991129-134258-0	Auckland-II	November 29, 1999	1:14:29
19991129-134258-1	Auckland-II	November 29, 1999	1:14:29
19991201-192548-0	Auckland-II	December 1, 1999	1:0:2
20010220-226-0	Auckland-IV	February 20, 2001	6:4:58
20010220-226-1	Auckland-IV	February 20, 2001	6:4:58
20010301-0310-0	Auckland-IV	March 1, 2001	9:14:49
20010301-0310-1	Auckland-IV	March 1, 2001	9:14:49
20010609-0613-0	Auckland-VI	June 9, 2001	4:6:0
20010609-0613-1	Auckland-VI	June 9, 2001	4:6:0
20010609-0613-e0	Auckland-VI	June 9, 2001	4:6:0
20010609-0613-e1	Auckland-VI	June 9, 2001	4:6:0
20020519-525	Bell-Lab-I	May 19, 2002	7:0:0

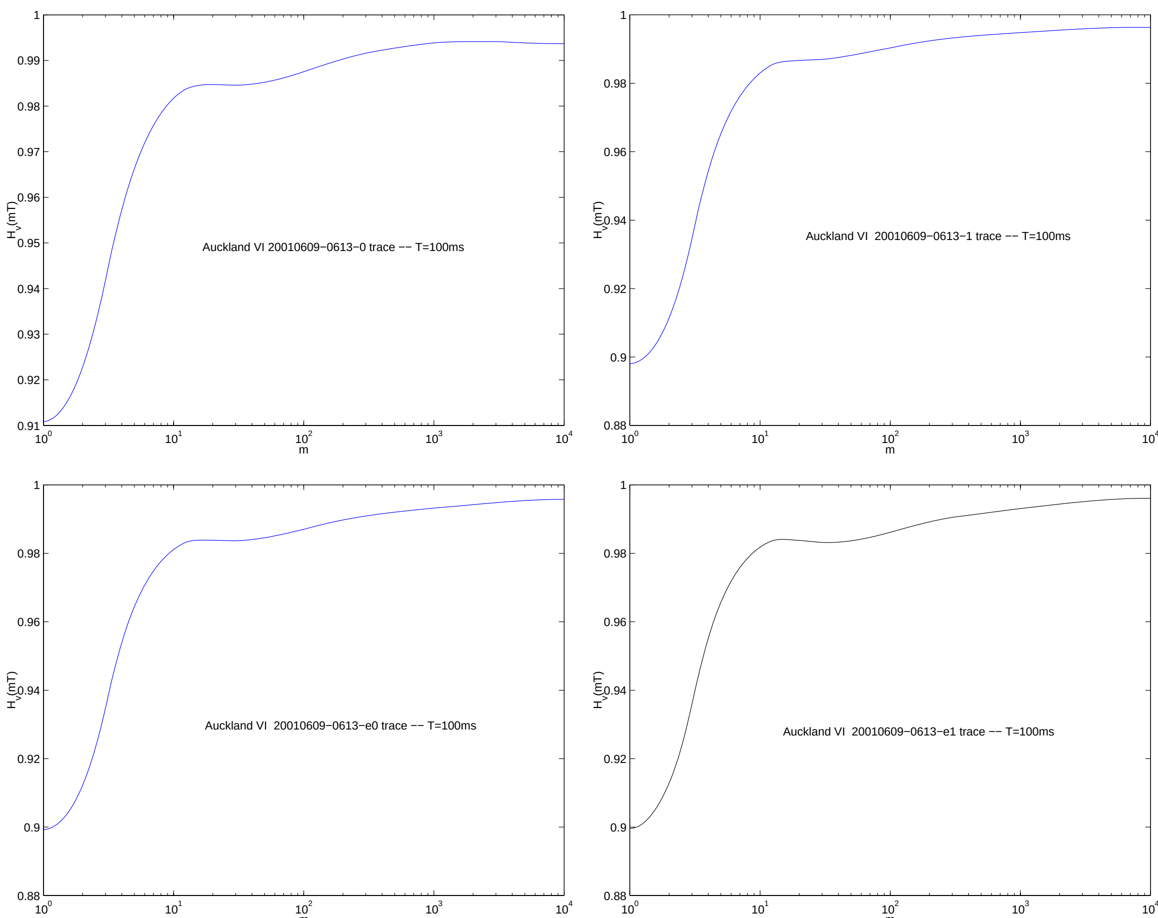

Fig. 7.   Estimated Index of Variability Curves for Auckland VI traces

[10] Riccardo Gusella, “Characterizing the Variability of Arrival Processes with Indexes of Dispersion”, IEEE *JSAC*, Vol. 9, No. 2, Feb. 1991.

[11] H. Heffes and D.M. Lucantoni, “A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance,” IEEE Journal on Selected Areas in Communications, Vol. SAC-4, No. 6, pp. 856-868, Sept. 1986.

[12] K. R. Krishnan, “A New Class of Performance Results for a Fractional Brownian Traffic Model,” Queueing Systems, Vol. 22, pp. 277-285, 1996.

[13] K. R. Krishnan, A. L. Neidhardt, A. Erramilli, “Scaling Analysis in Traffic Management of Self-Similar Processes,” Proc. 15th ITC, Washington, D.C., pp. 1087-1096, June 1997.

[14] G. Y. Lazarou, *On the Variability of Internet Traffic*, Ph.D. Dissertation, Electrical Engineering and Computer Science, University of Kansas, 2000.

[15] W. Leland, M. Taqqu, W. Willinger, D. Wilson, ”On the Self-Similar Nature of Ethernet Traffic (Extended Version),” IEEE/ACM Trans. on Net., Vol.2, No.1, Feb. 1994.

[16] W. E. Leland, D. V. Wilson, ”High Time-Resolution Measurement and Analysis of LAN Traffic: Implication for LAN Interconnection,” INFOCOM’91, Vol. 3, pp. 11d.3.1, 1991.

[17] A. L. Neidhardt, J. L. Wang, “The Concept of Relevant Time Scales and Its Application to Queueing Analysis of Self-Similar Traffic (or Is Hurst Naughty or Nice?),” Bellcore, Red Bank, New Jersey, September 12, 1997, {arnie, jwang}@bellcore.com.

[18] I. Norros, ”A storage model with self-similar input,” Queueing Syst. Vol. 16. pp. 387-396, 1994.

[19] K. Park, W. Willinger, Editors, *Self-Similar Network Traffic And Performance Evaluation*, John Wiley & Sons, 2000.

[20] V. Paxson, S. Floyd, ”Wide-Area Traffic: The Failure of Poisson Modeling,” IEEE/ACM Transactions on Networking, Vol. 3, No. 3, June 1995.

[21] M. S. Taqqu, V. Teverovsky, and W. Willinger, “Estimators for long-range dependence: an emperical study,” Appeared in “Fractals”, Vol. 3, No. 4, 1995.

[22] NLANR Network Traffic Packet Header (Long) Traces, [Online] Available: http://pma.nlanr.net/Traces/long/

[23] R.L. Burden, J.D. Faires, *Numerical Analysis* 7^{th} ed., Brooks/Cole, Thomson Learning, Inc., 2001.

[24] C.M. Reinsch, “Smoothing by Spline Functions,” *Numerische Mathematik*, Vol. 10, pp. 177-183, 1967.

[25] R.L. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York, New York, 1988.

[26] J. Fox, *Nonparametric Simple Regression: Smoothing Scatterplots,* Sage Publications, Thousand Oaks, California, 2000.

[27] B. Shahraray and D.J. Anderson, “Optimal Estimation of Contour Properties by Cross-Validated Regularization,” IEEE Transactions on Pattern Analysis and Machine Intelligense, Vol. 11, No. 6, June 1989.

[28] A. Horvath, M. Telek, “A Markovian Point Process Exhibiting Multifractal Behaviour and its Application to Traffic Modeling,” Fourth International Conference Matrix Analytic Methods in Stochastic Models, Adelaide, Australia, 14-18 July.

[29] Q. Li, D. Mills, “Investigating the Scaling Behavior, Crossover and Anti-persistence of Internet Packet Delay Dynamics,” Globecom’99, Vol. 3, pp. 1843-1852, 1999.

[30] K. Kant, “On aggregate Traffic Generation with Multifractal Properties,” Globecom’99 Vol. 2, pp.801-804, 1999.

[31] M.S. Taqque, Vadim Teverovsky, W. Willinger, “Is Network Traffic Self-Similar or Multifractal?” *Fractals*, 5 pp. 63-73.

[32] W. Willinger, V. Paxson, R.H. Riedi, M. Taqqu, “Long-Range Dependence and Data Network Traffic,” *Long-range Dependence: Theory and Applications*, P. Doukhan, G. Oppenheim, M.S. Taqqu, eds., Birkhauser, 2002.

[33] D. Veitch, P. Abry, “A Wavelet Based Joint Estimator for the Parameters of LRD,” *Special issue on Multiscale Statistical Signal Analysis and its Applications*, IEEE Trans. Info. Th., Vol. 45, No. 3, April 1999.

[34] V.J. Ribeiro, R.H. Reidi, R.G. Baraniuk, “Wavelets and Multifractals for Network Traffic Modeling and Inference,” ICASSP 2001.

[35] R.H. Riedi, M.S. Crouse, V.J. Ribeiro and R.G. Baraniuk, “A Multifractal Wavelet Model with Application to Network Traffic,” IEEE Trans. on Information Theory, Vol. 45, No. 3, April 1999.

[36] J. Levy Vehel, B. Sikdar, “A Multiplicative Multifractal for TCP Traffic,” Proceedings of Sixth IEEE Symposium on Computers and Communications, pp. 714-719, 2001.

[37] A.C. Gilbert, W. Willinger, and A. Feldmann, “Scaling Analysis of Conservative Cascades, with Applications to Network Traffic,” IEEE Trans. on Information Theory, Vol. 45, No. 3, April 1999.

[38] A. Feldman and W. Whitt, “Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models,” AT&T Laboratory – Research, December 1996.