

Analysis of LRD Series with Time-Varying Hurst Parameter

Análisis de Series LRD con Parámetro de Hurst Variante en el Tiempo

Sergio Ledesma Orozco, Gustavo Cerda Villafaña, Gabriel Aviña Cervantes,
Donato Hernández Fusilier and Miguel Torres Cisneros

Department of Electrical and Computer Engineering
University of Guanajuato
selo@salamanca.ugto.mx

Article received on September 05, 2008; accepted on January 19, 2009

Abstract

It has been previously shown that actual network traffic exhibits long-range dependence. The Hurst parameter captures the degree of long-range dependence; however, because of the nature of computer network traffic, the Hurst parameter may not remain constant over a long period of time. An iterative method to compute the value of the Hurst parameter as a function of time is presented and analyzed. Experimental results show that the proposed method provides a good estimation of the Hurst parameter as a function of time. Additionally, this method allows the detection on changes of the Hurst parameter for long data series. The proposed method is compared with traditional methods for Hurst parameter estimation. Actual and synthetic traffic traces are used to validate our results. The proposed method allows detecting the changing points on the Hurst parameter, and better results can be obtained when modeling self-similar series using several values of the Hurst parameter instead of only one for the entire series. A new graphical tool to analyze long-range dependent series is proposed. Because of the nature of this plot, it is called the *transition-variance* plot. This tool may be helpful to distinguish between LAN and WAN traffic. Finally, the software *LRD Lab** is deployed to analyze and synthesize long-range dependent series. The *LRD Lab* includes a simple interface to easily generate, analyze, visualize and save long-range dependent series.

Keywords: Estimation of Hurst parameter, self-similarity, long-range dependence, time-varying Hurst parameter.

Resumen

Ha sido previamente propuesto que el tráfico real de redes de computadoras exhibe dependencia de rango amplio. El parámetro de Hurst captura la cantidad de dependencia de rango amplio; sin embargo, debido a la naturaleza del tráfico en redes de computadoras, el parámetro de Hurst puede no permanecer constante durante un periodo largo de tiempo. Un método iterativo para calcular el valor del parámetro de Hurst como una función del tiempo es presentado y analizado. Los resultados experimentales demuestran que el método propuesto proporciona una buena estimación del parámetro de Hurst como una función del tiempo. Adicionalmente, este método permite la detección de cambios en el parámetro de Hurst para series largas. El método propuesto es comparado con métodos tradicionales para estimar el parámetro de Hurst. Series de datos reales y sintéticas son usadas para validar los resultados. El método propuesto permite detectar los puntos de cambio del parámetro de Hurst, y mejores resultados pueden ser obtenidos al modelar series similares a sí mismas usando varios valores del parámetro de Hurst en lugar de solamente uno para toda la serie. Una nueva herramienta gráfica para analizar series con dependencia de rango amplio es propuesta. Debido a la naturaleza de esta gráfica, ésta se llama gráfica de *transición de varianza*. Esta herramienta puede ser usada para distinguir entre tráfico LAN y WAN. Finalmente, el software *LRD Lab** es desarrollado para analizar y sintetizar series con dependencia de rango amplio. El *LRD Lab* incluye una interfase sencilla para generar, analizar, visualizar y almacenar series con dependencia de rango amplio.

Palabras clave: Estimación del parámetro de Hurst, similar así mismo, dependencia de rango amplio, parámetro de Hurst variante en el tiempo.

* The *LRD Lab* can be downloaded at <http://www.fimee.ugto.mx/profesores/sledesma/documentos/index.htm>

1 Introduction

A standard assumption of time series analysis is that observations separated by a large time span are roughly independent. However, in some time series, the observations are not independent and exhibit long-range dependence. Thus, long-range dependence involves the tail behavior of the autocorrelation function. The simplest models with long-range dependence are self-similar processes, which are characterized by a hyperbolically decreasing autocorrelation function. Self-similar and asymptotically self-similar processes are particularly attractive models because the long-range dependence can be characterized by a single parameter called the Hurst parameter, H (the degree of both self-similarity and long-range dependence increases as $H \rightarrow 1$.) On the other hand, it has been pointed out that it may be unrealistic to characterize the process only by a single value of H [Stoev et al., 2006].

Nowadays, computer network traffic affects significantly the performance of most Internet services [Krunz and Matta, 2002; Moulines, et al., 2002; Zukerman et al., 2003]. Practical network traffic modeling is crucial to efficiently accommodate data links, and may include the synthesis and analysis of long-range dependent series. Contrary to commonly held views that multiplexing traffic streams tends to produce smoothed-out aggregate traffic with reduced burstiness, aggregating self-similar traffic streams can actually intensify burstiness rather than diminish it [Leland, et al., 1994]. It is demonstrated in [Erramilli, et al., 1996; Karagiannis, et al., 2004; Ramirez and Torres, 2006] that long-range dependence is not merely relevant for queuing performance but it is a dominant characteristic for determining several issues of network engineering.

On the other hand, the synthesis of long-range dependent series has been intensively analyzed [Paxson, 1997; Ledesma et al., 2000; Purczynski and Wlodarski, 2005; Ledesma et al., 2007]. Several methods to estimate the Hurst parameter have been previously proposed, [Beran et al., 1995]. However, the detection on changes of the Hurst parameter is an open area for research. This paper is organized as follows. Section 2 lists some of the traditional methods to estimate the Hurst parameter. Section 3 presents a new method to detect changes on the Hurst parameter. Section 4 shows experimental results to validate the proposed method. Finally, section 5 offers conclusions and provides direction for future work.

2 Estimation of H

Consider a stochastic process $X = \{X_i; i = 0, 1, 2, \dots\}$ where all $\{X_i\}$ have a common mean μ , a common finite variance σ^2 and a stationary autocorrelation function $r(k)$ that depends only on k , but not on i . A process with these properties is called second order or wide-sense stationary.

Now, for each $m = 1, 2, 3, \dots$, let

$$X^{(m)} = \{X_k^{(m)}; k = 0, 1, 2, \dots\},$$

denote a new time series obtained by averaging the original series X ; that is,

$$X_k^{(m)} = \frac{X_{km-m} + \dots + X_{km-1}}{m}. \quad (1)$$

Let σ_m^2 and $r^{(m)}$ denote the variance and the autocorrelation function of $\{X_k^{(m)}\}$, then X is called an exactly second-order self-similar process, if, for all $m = 1, 2, 3, \dots$

$$\text{Var}(X^{(m)}) = \sigma^2 m^{-(2-2H)} \quad \text{and} \quad r^{(m)}(k) = r(k), \quad k \geq 0.$$

Additionally, X is called an asymptotically second-order self-similar process, if, for all k large enough,

$$r^{(m)}(k) \rightarrow r(k), \text{ as } m \rightarrow \infty .$$

Additional information about second-order exact self-similar process and second-order asymptotical self-similar process can be found in [Tsybakov and Georganas, 1998].

Before ending this section, it is important to mention that the new sequence $X_k^{(m)}$ is frequently used to estimate the Hurst parameter [Roughan, et al., 1998], and that generally, the Hurst parameter can be estimated using the:

- Analysis of the variances of the aggregate processes $X^{(m)}$.
- Time-domain analysis based on the re-scaled adjusted range (R/S-statistics).
- Frequency domain analysis based on the periodogram or Whittle's estimator.
- Wavelet-based estimator, see Veitch and Abry (1998, 1999).

2.1 The Aggregate Processes $X^{(m)}$

A valuable tool for assessing burstiness over different time-scales is the *variance-time* plot [Paxson and Floyd, 1995]. For second-order self-similar processes, the variances of the aggregate processes $X^{(m)}$, decrease linearly (for large m) in log-log plots against m with slopes arbitrarily flatter than -1. On the other hand, the variances of $X^{(m)}$ for short-range dependent processes will eventually decrease linearly in log-log plots against m with a slope equal to -1. A variance-time plot is obtained by plotting $\log[\text{Var}(X^{(m)})]$ against $\log(m)$ and is used to determine whether the data support the asymptotic behavior of the form

$$\text{Var}[X^{(m)}] \propto am^{-(2-2H)} \tag{2}$$

$m \rightarrow \infty$ when $a > 0$ and $1/2 < H < 1$. Note that in this case a is a positive constant to indicate that the variance of $X^{(m)}$ is proportional to $m^{-(2-2H)}$. Thus, an estimate of the exponent of m , $-(2-2H)$, can be obtained from the slope of a linear regression fitted through the resulting points in the plane, ignoring small values of m .

2.2 R/S-Statistics

Let $X = \{X_k; k = 1, 2, \dots, N\}$ be a record containing N readings uniformly spaced in time from $k = 1$ to $k = N$, and let

$$X_k^* = \sum_{u=1}^k X_u$$

and

$$W(k, n, u) = X_{k+u}^* - X_k^* - \frac{u}{n} [X_{k+n}^* - X_k^*].$$

Thus, $n^{-1}X_n^*$ is the average of the first n readings and $n^{-1}[X_{k+n}^* - X_k^*]$ is the average of readings within the sub-record from time $k + 1$ to time $k + n$ [Mandelbrot, 1969].

$S^2(k, n)$ is defined as the sample variance of the sub-record from time $k + 1$ to time $k + n$, namely,

$$S^2(k, n) = \frac{1}{n} \sum_{u=k+1}^{k+n} X_u^2 - \left[\frac{1}{n} \sum_{u=k+1}^{k+n} X_u \right]^2$$

$R(k, n)$ is defined as follows

$$R(k, n) = \max_{0 \leq u \leq n} W(k, n, u) - \min_{0 \leq u \leq n} W(k, n, u).$$

The objective of the R/S analysis of an empirical record is to infer the degree of self-similarity, H , for the process that presumably generated the record under consideration. Hurst's study of the re-scaled adjusted range found that many historical records appeared to be well represented by

$$E \left[\frac{R(k, n)}{S(k, n)} \right] \propto cn^H, \text{ as } n \rightarrow \infty, c > 0. \quad (3)$$

The ratio $R(k, n)/S(k, n)$ is then called a re-scaled adjusted range. To construct a box plot of R/S , a sequence of logarithmically spaced values of n is selected. For each n , a number of starting points k is selected. Plotting $\log \frac{R(n)}{S(n)}$ versus $\log(n)$ results in the *re-scaled adjusted range* plot.

2.3 The Periodogram

Parameter estimation for random fields and random processes based on smoothed periodogram has a long history. The idea derives from suggestive forms of Whittle's estimation procedure, which has formed the backbone of asymptotic estimation since its discovery [Heyde and Gay, 1995]. Thus, instead of using the autocorrelation function to estimate H , a mathematically equivalent expression in the frequency domain may be used [Michiel and Laevens, 1997]. The resulting quantity is called the power spectrum or power spectral density [Oppenheim and Schaffer, 1989].

Specifically, this method is based on the behavior of the spectral density $f(\lambda)$ at the origin as shown

$$f(\lambda) = \frac{1}{\sigma^2} \sum_k e^{jk\lambda} r(k) \propto L(\lambda) |\lambda|^{-(2H-1)}, \text{ as } |\lambda \rightarrow 0| \quad (4)$$

where $r(k)$ is the autocorrelation function, $L(\lambda)$ is a slowly varying function and $1/2 < H < 1$. The absence of any limit law makes the statistics, corresponding to the R/S analysis or the *variance-time* plot inadequate. A more refined data analysis is possible for maximum likelihood-type estimates (MLE) and related methods based on the periodogram $I_X(\cdot)$ of X , defined by

$$I_X(\lambda) = \frac{1}{2\pi N} \left| \sum_{k=1}^N X_k e^{jk\lambda} \right|^2, \quad j = \sqrt{-1}.$$

3 Proposed Method

The analysis of real traffic in [Leland, et al., 1994] has shown that the Hurst parameter can be expected to change during a measurement period. It was also concluded in [Leland, et al., 1994] that modeling the changing points of H may be needed in the future in order to produce more realistic traffic models. Moreover, it has been established in [Paxson and Floyd, 1995] that while large-scale correlations are present in wide area network traffic traces, it might be difficult to characterize the correlations over the entire trace with a single Hurst parameter. Clearly, further research is required to understand the correlation structure of network traffic. In this section, an algorithm for

detecting the changing points of H is developed and evaluated. Additionally, it is shown how this algorithm can be used to obtain a fast estimation of the Hurst parameter.

3.1 Analysis of Self-Similar Traffic with Variable Hurst Parameter

Apart from making parameter estimation for large self-similar series computationally feasible, the procedure proposed by [Beran and Terrin, 1992] provides a method for checking whether H (and possibly additional parameters) remains constant over time series. The issue of deciding whether there are any deviations or not in H over time can be a complex problem. These deviations could be the result of actual changes in the dependence structure in the data or due to randomness. Additionally, estimates of H turn out to vary considerably when they are calculated from disjoint parts of a long-range dependent series. In order to assess quantitatively how much the estimates of H can vary when they are estimated from different portions of the data, Beran and Terrin (1992) obtained the joint asymptotic distribution of the Whittle estimates of H based on disjoint sub-series.

In order to detect the changing points of the Hurst parameter in a time series, the asymptotic relations of equations (2), (3), and (4) can be used. Equations (2) and (3) provide a natural way to detect changes in the value of the Hurst parameter as a function of time, because they are time domain asymptotic relations of long-range dependence (LRD). Equation (2) is the time domain counterpart of equation (4). Although equation (4) can be used to detect the changing points in the value of H , an iterative procedure based on this equation cannot be developed. Additionally, using the periodogram to detect the changing points in the Hurst parameter has already been analyzed by [Beran and Terrin, 1992]. Finally, note that equation (2) is the simplest method to estimate H , and that it can be iteratively implemented, which is one of the main advantages of using this approach to analyze data series with variable Hurst parameter.

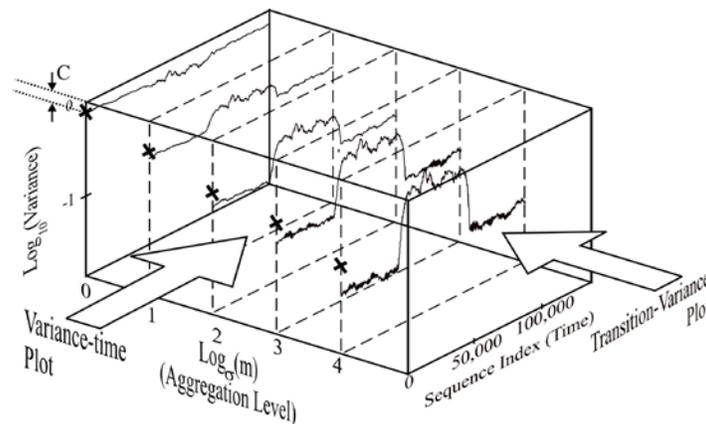


Fig. 1. Relationship between the variance-time plot and the transition-variance plot

The analysis of self-similar traffic using the variance-time plot can be extended in a natural way to detect the changing points in the value of H . From equation (2), it can be seen that the Hurst parameter may be calculated from the exponent of m . Thus, in order to get an estimate of H , the variances of the aggregate processes $X^{(m)}$, see equation (14), are plotted against m in log-log scales as described in section 2.1. Figure 1 shows how the variance-time plot is constructed; from this figure it can be clearly seen that the variance-time plot does not provide information about how the Hurst parameter is changing over time.

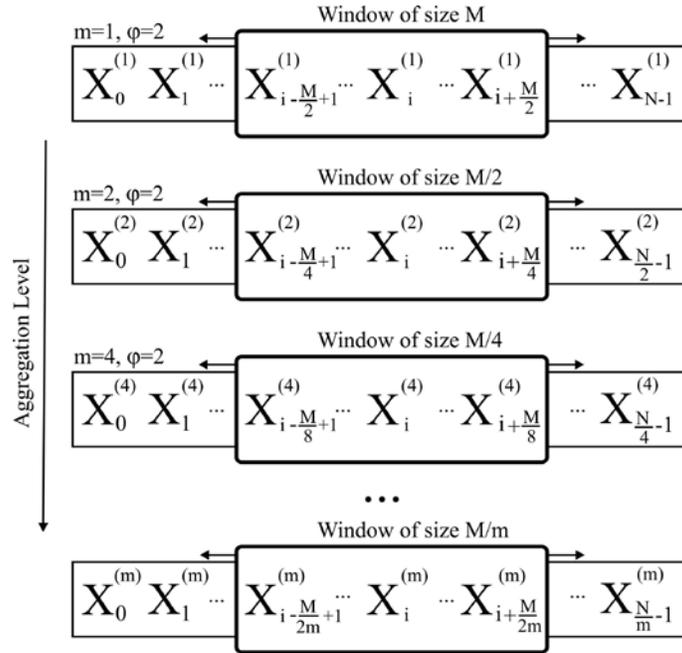


Fig. 2. Sliding window for several aggregation levels

To detect changes on the Hurst parameter consider the following discussion. First, an aggregation base, φ , must be defined, see Figure 2. This results on logarithmic spaced values for m as shown

$$m = \varphi^0, \varphi^1, \varphi^2, \varphi^3, \dots \tag{5}$$

Typical values for φ are: 2 or 3; bigger values for φ are recommended only for very long traces because the sequence length reduces quickly when the aggregation process is repeatedly applied. The value of φ must be adjusted to obtain the desired resolution, a value of 2 results on the following aggregation values of m : 1, 2, 4, 8, ...; a value of 3, will produce the aggregation levels of m : 1, 3, 9, 27, It is important to mention that as m (the aggregation level) becomes larger and larger, the variance is computed using fewer and fewer points. Therefore, there is a maximum value of m for which, it does not make any sense to keep computing the variance (as it will be small).

Consider now, a window of size M whose elements are taken from a self-similar series X of length N with $M \ll N$, see Figure 2. Hence,

$$\eta = \frac{N}{m}$$

is the length of $X^{(m)}$. Thus, an estimate of the sample variance for an aggregation level m using only the elements of the window can be calculated as follows:

$$\overline{Var}_i(X^{(m)}) = \frac{mQ_i^{(m)}}{M-m} - \frac{[mR_i^{(m)}]^2}{M(M-m)}, \tag{6}$$

where

$$Q_i^{(m)} = \sum_{j=i-\frac{M}{2m}+1}^{i+\frac{M}{2m}} [X_j^{(m)}]^2 \tag{7}$$

and

$$R_i^{(m)} = \sum_{j=i-\frac{M}{2m}+1}^{i+\frac{M}{2m}} X_j^{(m)}, \tag{8}$$

for $i = 0, 1, 2, \dots, \eta - 1$. Observe that $R_i^{(m)}$ can be used to compute an estimate of the average value of $X^{(m)}$ while $Q_i^{(m)}$ can be used to compute an estimate of the average value of $(X^{(m)})^2$.

By sliding the window through the whole series, it is possible to see how the variance changes through time. Moreover, the values of $Q_i^{(m)}$ and $R_i^{(m)}$ can be iterated as:

$$Q_{i+1}^{(m)} = Q_i^{(m)} + [X_{i+1+\frac{M}{2m}}^{(m)}]^2 - [X_{i+1-\frac{M}{2m}}^{(m)}]^2$$

And

$$R_{i+1}^{(m)} = R_i^{(m)} + X_{i+1+\frac{M}{2m}}^{(m)} - X_{i+1-\frac{M}{2m}}^{(m)}$$

Thus, to see if the value of H remains constant through the whole series, we propose to plot $y_i(m)$, using equation (9), for several values of m and i ,

$$y_i(m) = \log \left[\frac{\overline{\text{Var}_i(X^{(m)})}}{\overline{\text{Var}_i(X^{(1)})}} \right]. \tag{9}$$

Due to the nature of this plot, we have called *transition-variance* plot to the resulting graph. Figure 1 shows the relationship between the *variance-time* plot and the *transition-variance* plot. It is seen from this figure that the *variance-time* plot provides an estimate of H for the whole series, while the *transition-variance* plot indicates how the Hurst parameter is changing with time. A typical *variance-time* plot can be seen from the left view of Figure 1. Mathematically, the *transition-variance* plot can be analyzed by writing equation (5) as

$$m = \varphi^\mu, \quad \mu = 0, 1, 2, 3, \dots \tag{10}$$

substituting equation (10) in equation (2), we obtain

$$\text{Var}[X^{(m)}] \propto a\varphi^{-\mu(2-2H)}. \tag{11}$$

Applying the logarithm function on both sides of equation (11), we get

$$\log[\text{Var}(X^{(m)})] = -\mu(2 - 2H)\log(\varphi) + C.$$

A transition-variance plot is built by plotting $\log[\text{Var}(X^m)]$ for different values of μ ; notice that for $\mu = 0$,

$$\log[\text{Var}(X^{(m)})] = C,$$

therefore, C is an inconsequential constant that determines the absolute position of the lines of a *transition-variance* plot, see Figure 1. This constant is irrelevant to this analysis because it is not a function of H . In order to mathematically interpret a *transition-variance* plot, it is important to see how $\log[\text{Var}(X^{(m)})]$ changes when m changes, namely when μ changes. This can be observed by computing the partial derivative of $\log[\text{Var}(X^{(m)})]$ with respect to μ as shown

$$z_\mu = \frac{\partial \log[\text{Var}(X^{(m)})]}{\partial \mu} = (2H - 2)\log(\varphi), \tag{12}$$

where z_μ indicates the line spacing on a *transition-variance* plot for two consecutive values of μ . Because of the self-similar structure of the series under consideration, the values of z_μ must be approximately the same no matter what value of μ or φ is chosen. Any irregularity on the spacing is a clear symptom that the process originating the series under analysis is not an exact self-similar process. More over, the variance of z_μ can be clearly used to establish how close the data under analysis is to the one generated from an exact self-similar process.

Additionally, an aggregation process over the elements of the window can be performed to estimate

$$\text{Var}(X^{(m)}).$$

To speed up the computation of the *transition variance-time* plot, we suggest an iterative algorithm to compute the aggregation process described by

$$X_k^{((i+1)\varphi)} = \{X_k^{((i+1)\varphi)}, k = 0, 1, 2, \dots\} \tag{13}$$

for $i = 1, 2, 3, \dots$, where $X_k^{((i+1)\varphi)}$ denotes a new time series obtained by averaging $X_k^{(i\varphi)}$; that is,

$$X_k^{((i+1)\varphi)} = \frac{X_{k\varphi-\varphi}^{(i\varphi)} + \dots + X_{k\varphi-1}^{(i\varphi)}}{\varphi}. \tag{14}$$

Specifically, the aggregating process $X_k^{((i+1)\varphi)}$ can be computed from $X_k^{(i\varphi)}$. This means that for $\varphi = 2$, $X^{(4)}$ can be calculated by aggregating $X^{(2)}$, $X^{(8)}$ from $X^{(4)}$, and so on, making the computation of the aggregated process $X^{(m)}$ fast and easy.

4 Results

We proceed now to analyze synthetic and real LRD series with variable Hurst parameter using the proposed method.

4.1 Transition Detection

There are several existing methods to generate series that exhibit LRD, see [Ledesma, et al., 2007]. In particular, the experiments of this section were performed using the fractional Gaussian noise which is one of the simplest models that exhibit self-similarity. Thus, a synthetic self-similar series of length 131,072 was created using the algorithm proposed in [Ledesma, et al., 2007]; the first third part of the series was synthesized using a value of $H=0.55$, while the remaining part was synthesized using a value of $H=0.85$. Figure 3 shows the *transition-variance* plot for this series. It is observed from this figure that the *transition-variance* plot is very helpful for detecting the changing points of H . It is important to mention that the window size affects the resolution of the plot; as the window size increases, the *transition-variance* plot exhibits a more stable behavior and random transitions on the Hurst parameter are ruled out, this can be easily observed by comparing Figure 3(a) and 3(b). Note that wider windows require longer series.

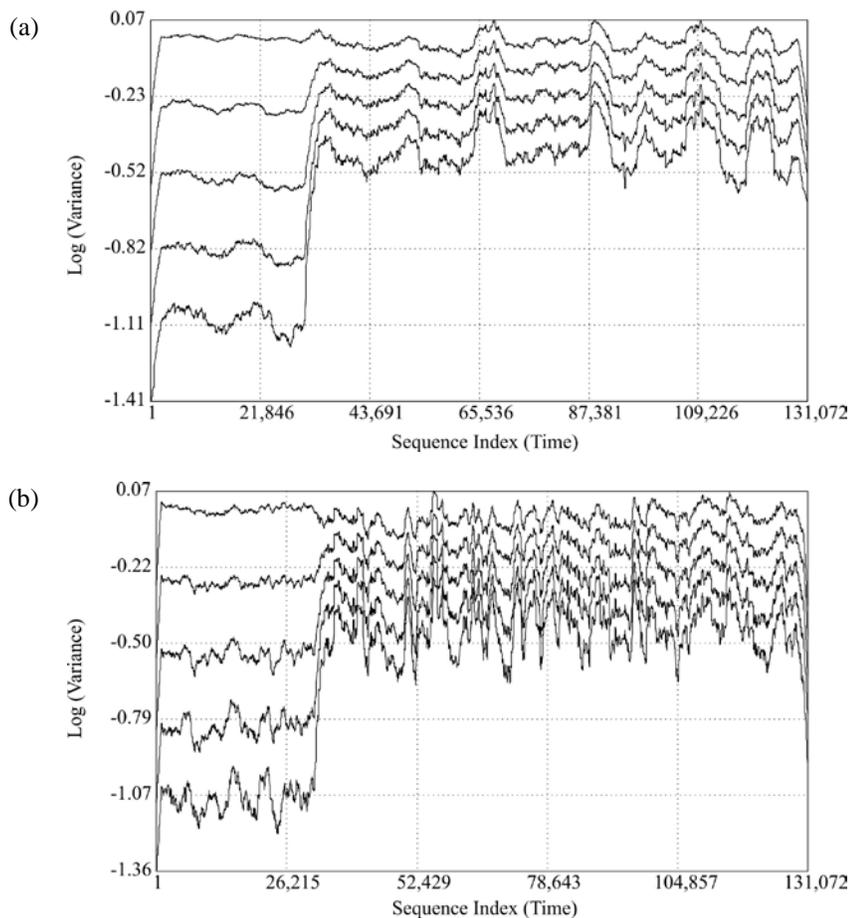


Fig. 3. Transition-variance plot for a self-similar sequence with $H=0.55$ and $H=0.85$.
(a) Sliding window size = 4,096; (b) Sliding window size = 2,048

On the other hand and despite the fact that a Hurst parameter transition can be easily inferred from Figures 3(a) and 3(b), it is important to note that the Hurst parameter is expressed by the line spacing of these figures, and not by the vertical distance from the lines to the *Time-axis*. Therefore, it would be possible to detect a rough estimate of the Hurst parameter by plotting only $\log(\text{Variance})$ for two consecutive values of m , namely 2 and 4 (or 4 and 8).

However, a more stable estimate for H can be computed using more aggregation values for m as more information from the series self-similar structure is used to estimate H ; after all the line spacing should be approximately the same for two consecutive values of m . If the line spacing on a *transition-variance* plot is not uniform this would mean an irregularity on the series self-similar structure; namely the process is only asymptotically self-similar. This can be mathematically understood by the fact that when the variance of an exact self-similar series is plotted for several aggregation levels (several values of m) using logarithmic scales, a straight line should be constructed; a straight line means that the $\log(\text{Variance})$ is a linear function of $\log(m)$. Thus, any deviation from linearity means that the data under analysis is not produced by an exact self-similar process.

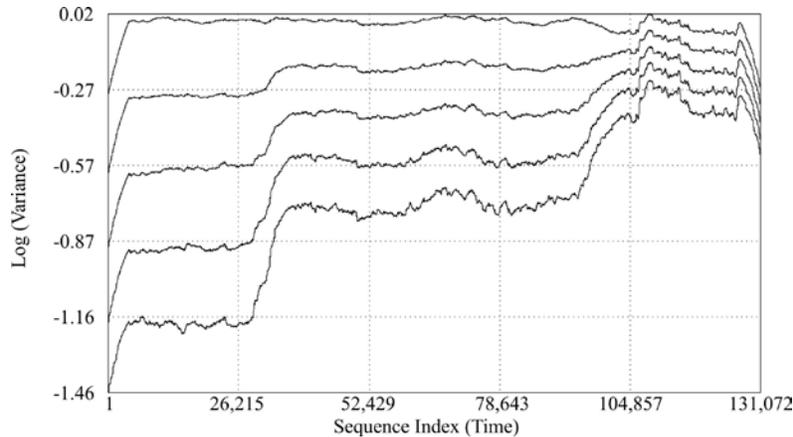


Fig. 4. *Transition-variance* plot for a self-similar sequence with $H=0.51$, $H=0.70$ and $H=0.90$ (Sliding window size = 8,192)

Before analyzing real series, observe Figure 4 which shows the *transition-variance* plot for a synthetic sequence with three different Hurst parameters. The first third of the sequence was created with a target value of $H=0.51$, the second third using $H=0.70$ and the last part of the sequence using $H=0.90$; observe the space between lines: as the value of H increases the spacing decreases.

Real Ethernet traffic traces will be now analyzed. These traces were obtained from an Internet traffic archive (<http://ita.ee.lbl.gov/html/traces.html>). A brief discussion of some of the traces of LAN and WAN traffic in this archive is presented next.

The BC traces contain a million packet arrivals seen on an Ethernet network at the Bellcore Morristown research and engineering facility, [Leland and Wilson, 1991; Leland, et al., 1994]. Two of the traces in this archive are LAN traffic (with a small portion of transit WAN traffic), and the other two are WAN traffic.

The trace BC-pOct89 began at 11:00 on October 5, 1989, and ran for about 11,760 seconds. A *transition-variance* plot is shown in Figure 5(a). We can observe from this figure that the Hurst parameter is almost constant for the first half part of the sequence. Then, a slight transition on the Hurst parameter can be appreciated at approximately the middle of the sequence; a bigger line spacing means a decrement on the value of H .

The trace BC-Oct89Ext began at 23:46 on October 3, 1989, and captured the first 1 million external arrivals (packets headed between Bellcore and the rest of the Internet), ending about 122,797 seconds later. The trace BC-Oct89Ext4 comes from the 4th tape of a 307-hour trace begun at 14:37 on October 10, 1989. The tape started at time-stamp 774,019, about 215 hours into the trace, and BC-Oct89Ext4 ends about 75,943 seconds later. Figure 5(b) shows the *transition-variance* plot for the series BC-Oct89Ext4. From Figure 5(b), it can be observed that WAN traffic has severe changes on the Hurst parameter when compared to LAN traffic (Figure 5(a)). Moreover, the *transition-variance* plot from LAN and WAN traffic are clearly different.

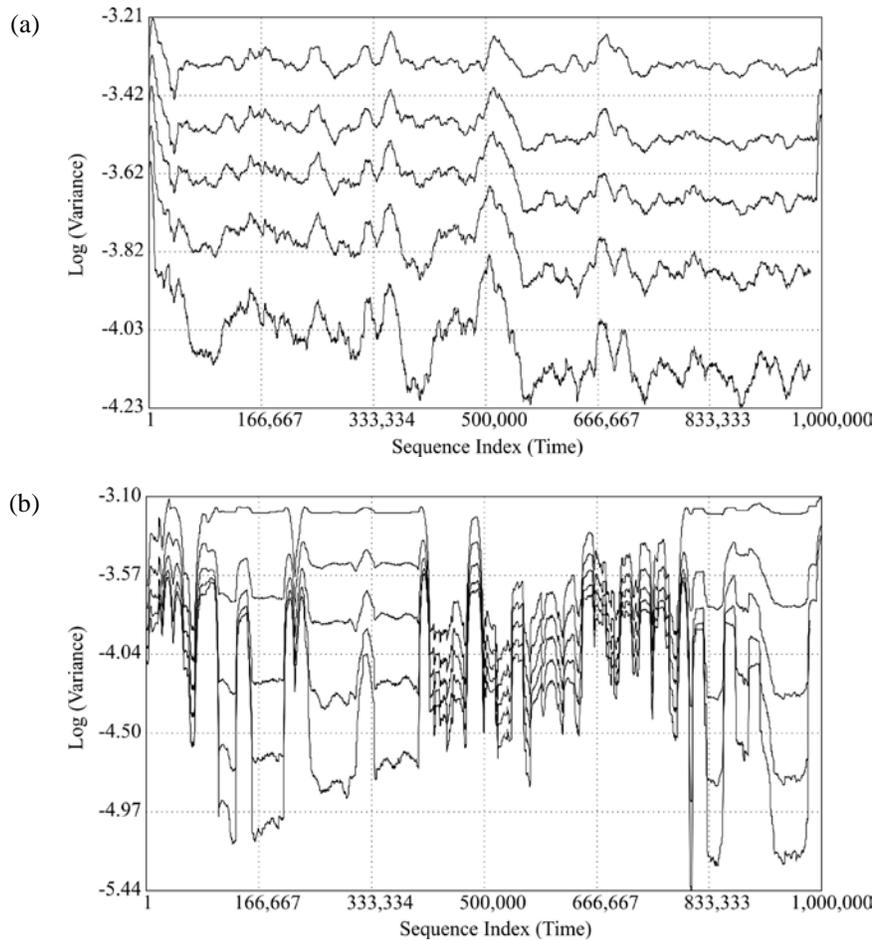


Fig. 5. Transition-variance plot with a sliding window size of 16,396
 (a) LAN traffic, trace BC-pOct89; (b) WAN traffic, Trace BC-Oct89Ext4

4.2 Fast Estimation of H

Most estimates and test statistics have a slower rate of convergence for long-range dependent processes. Thus, assuming independence or some kind of short-range dependence will lead to underrating uncertainty (measured by the size of confidence intervals) by a factor which approaches infinity as the sample size approaches infinity, [Beran and Terrin, 1992b]. The rate of convergence for many standard statistics is determined by the value of the parameter H . For reliable statistical inference, it is therefore important to obtain a good estimate of H from the data [Fox and Taqqu, 1985].

From a statistical point of view, LRD may, in some cases, have unexpected and perhaps serious consequences. For example, the accuracy of a statistical measurement generally depends on having a large enough sample for the statistics to converge meaningfully. Confidence intervals (CI) are used widely in performance analysis to gauge the accuracy of parameter estimates. The conventional CI calculation not only assumes that measurement errors are Normally distributed, but also that they are i.i.d. For short-range dependent processes (i.e., having an exponentially decreasing autocorrelation function), the correlations become negligible after a finite and usually small lag, and confidence intervals are reasonably accurate. For long-range dependent processes, however, this is not the case [Garret and Willinger, 1994]. The importance of a fast estimation of H lies in the fact that it provides information, which can help the routing and admission control devices to improve the use of the resources.

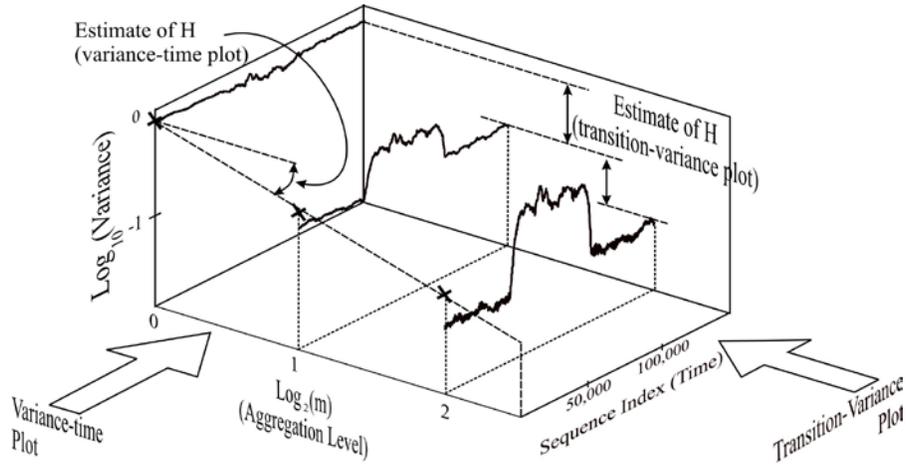


Fig. 6. Estimation of H from the *variance-time* plot and the *transition-variance* plot

An estimate of the Hurst parameter can be obtained from the *transition-variance* plot as shown in Figure 6. Consider equation (2), which can be written as:

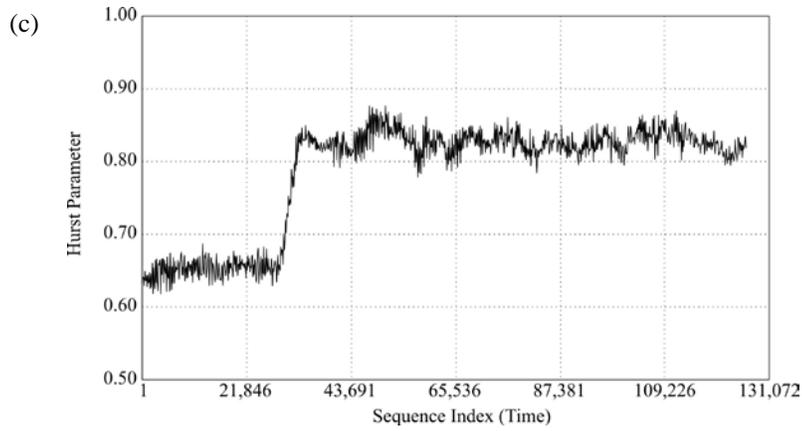
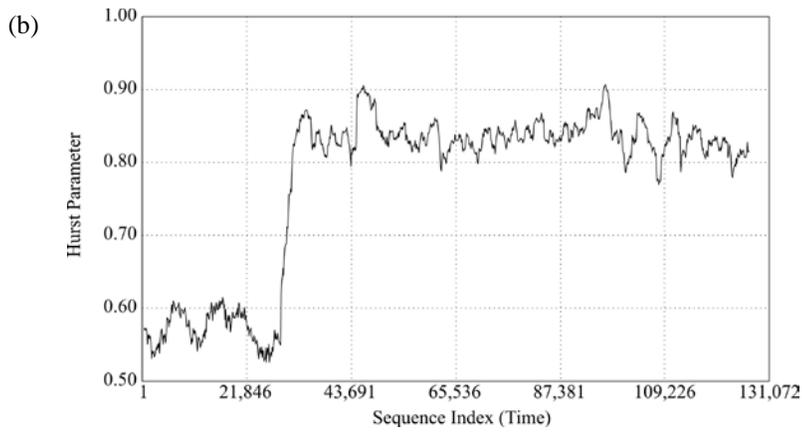
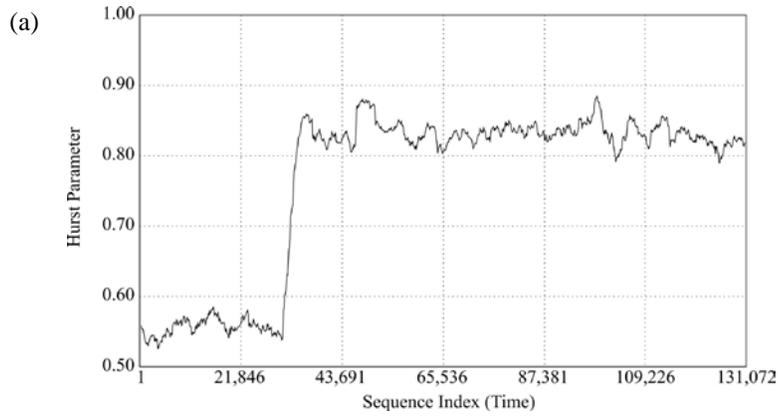
$$\log[\text{Var}(X^{(m)})] = (2H - 2)\log(m) + C.$$

Now, consider two consecutive aggregation levels, $m_j = \varphi^j$ and $m_{j+1} = \varphi^{j+1}$, then, an estimate for the Hurst parameter from this two aggregation levels can be calculated as

$$\begin{aligned} \overline{H}_{i,j} &= 1 - \frac{1}{2} \frac{\log[\text{Var}_i(X^{(m_j)})] - \log[\text{Var}_i(X^{(m_{j+1})})]}{\log(m_{j+1}) - \log(m_j)} \\ &= 1 - \frac{1}{2\log(\varphi)} \left[\log[\text{Var}_i(X^{(m_j)})] - \log[\text{Var}_i(X^{(m_{j+1})})] \right]. \end{aligned}$$

Thus, a Hurst parameter estimate for each value of i can be calculated when there are K aggregation levels by

$$\overline{H}_i = \frac{1}{K-1} \sum_{j=1}^{K-1} \overline{H}_{i,j}, \quad 0 \leq i < N. \tag{15}$$



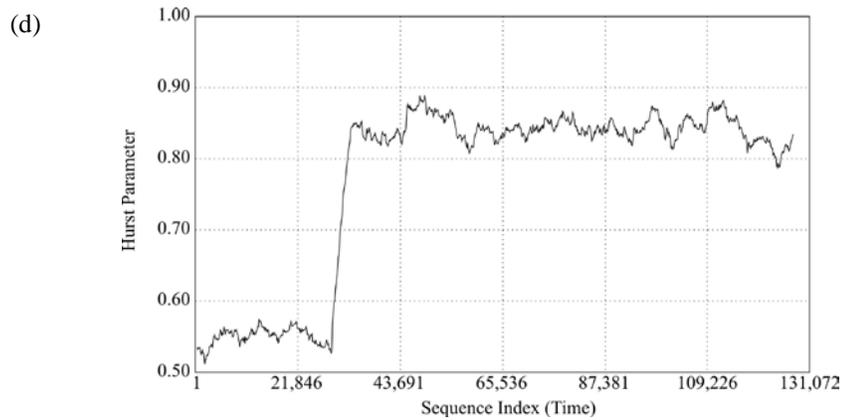


Fig. 7. Hurst-time plot for a synthetic self-similar series with $H=0.55$ and $H=0.85$
 (a) Proposed method; (b) *Variance-time* plot; (c) R/S plot; (d) Whittle's estimator

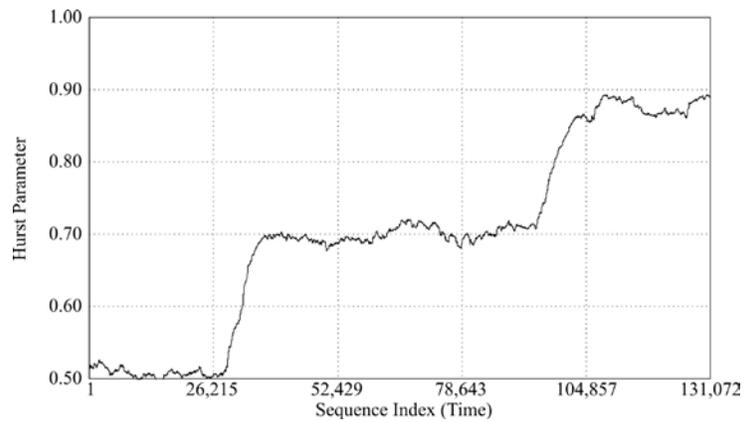


Fig. 8. Hurst-time plot for a synthetic self-similar series with $H=0.51$, $H=0.70$ and $H=0.90$
 (Sliding window size=8,192)

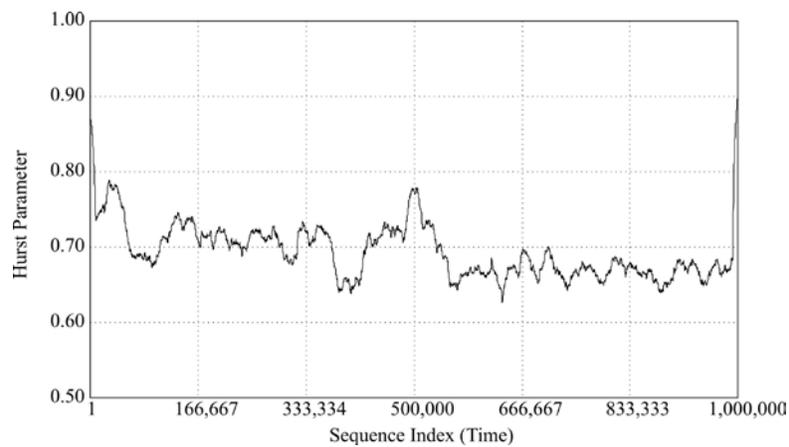


Fig. 9. Hurst-time plot for the series BC-pOct89

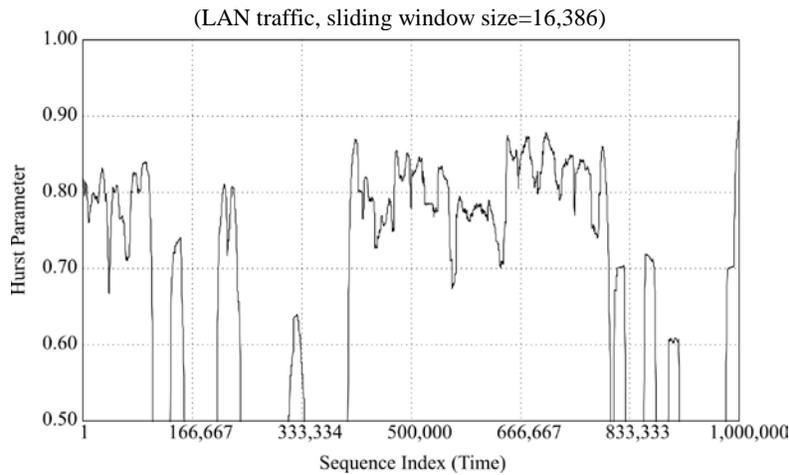


Fig.10. Hurst-time plot for the series BC-pOct89Ext4 (WAN traffic, sliding window size=16,386)

Figure 6 illustrates graphically how to estimate the Hurst parameter from the *variance-time* plot and the *transition-variance* plot. The proposed method computes the *transition-variance* plot and estimates the Hurst parameter continuously. In other words, the *transition-variance* plot may be iteratively implemented to plot the Hurst parameter as a function of time. We called this plot the Hurst-time plot. Figure 7 shows the Hurst-time plot using the proposed method and three traditional methods (with an overlapping window size of 4,096). Figures 8, 9, and 10, show the respective Hurst-time plot for the synthetic and real LRD series previously discussed using the proposed method. A closer look to Figure 10 reveals that WAN traffic exhibits periods of self-similarity ($H > 0.5$), as well as periods of non self-similarity ($0 < H \leq 0.5$). By comparing Figure 9 and 10, it is clear that WAN traffic exhibits more changes in the Hurst parameter than LAN traffic.

Table 1. Running Time to compute the Hurst-time plot using an overlapping window on a series of size 131,072

Method	Running Time
Proposed method	0.008 seconds
Variance-time plot	60.20 seconds
R/S plot	98.69 seconds
Whittle's estimator	2 hours and 20 minutes

Table 1 shows the required running time to calculate the plots of Figure 7 using the proposed method, the *variance-time* plot, the R/S plot, and Whittle's plot on a computer with an Intel Xeon 3.2 GHz processor. The Hurst parameter was calculated using an overlapping sliding window of size 4,096. As it can be seen the proposed method is faster than the other ones; while Whittle's estimator is the slowest.

Finally, we note that the estimation of H from the *transition-variance* plot underestimates the value of the Hurst parameter for highly correlated series (values of H close to 1). This underestimation of H is also presented in the *variance-time* plot as it is well known, see Karagiannis (2004).

5 Conclusions

A new method to detect changes on the Hurst parameter for self-similar network traffic, called *transition-variance* plot, is proposed. This is a time-based estimation method to assess whether or not the Hurst parameter remains constant over a period of time. It detects the changing points on the Hurst parameter. Thus, this method allows the modeling of LRD series using several values of the Hurst parameter instead of only one for the entire series. It is

established that the *transition-variance* plot is a great descriptor for analysis of self-similar traffic. It allows distinguishing between LAN and WAN traffic.

An iterative algorithm is suggested to speed up the computation of the *transition-variance* plot. Several experimental results using synthetic and real self-similar traces are performed to validate our results. Finally, an iterative procedure to quickly estimate the Hurst parameter is reviewed. This iterative procedure generates a new plot called Hurst-time plot and allows observing the value H as a function of time.

The software LRD Lab is presented. This software is a graphical tool to create and analyze long-range dependent series; it generates the *variance-time* plot, the R/S pox plot, the periodogram and Whittle's estimator. Moreover, this software includes the two new plots: *transition-variance* plot and Hurst-time plot.

It would be interesting to explore the use of artificial intelligence algorithms to the analysis of network traffic that exhibits LRD.

Acknowledgments

This work was sponsored by CONACYT, DINPO and PROMEP.

6 References

- 1 **Beran, J. (1992).** Statistical methods for data with long-range dependence. *Statistical Science*, 7(4), 404-427.
- 2 **Beran, J., Sherman, R., Taqqu, M. S. & Willinger, W. (1995).** Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43(234), 1566-1579.
- 3 **Beran, J. & Terrin, N. (1992).** Estimation of the long-memory parameter, based on a multivariate central limit theorem. *Journal of Time Series Analysis*, 15(3), 269-278.
- 4 **Cappe, O., Moulines, E., Pesquet, J. C., Petropulu A. P. & Yang, X. (2002).** Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Processing Magazine*, 19(3), 14-27.
- 5 **Erramilli, A., Narayan, O. & Willinger, W. (1996).** Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(29), 209-223.
- 6 **Fox, R. & Taqqu, M. S. (1985).** Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics*, 4(2), 517-532.
- 7 **Garret, M. W. & Willinger, W. (1994).** Analysis, modeling and generation of self-similar VBR video traffic. *Computer Communications Review*, 24(4), 269-280.
- 8 **Heyde, C. C. & Gay, R. (1993).** Smoothed periodogram asymptotics and estimation for processes and fields with possible long-range dependence. *Stochastic Processes and Their Applications*, 45(1), 169-182.
- 9 **Karagiannis, T., Molle, M. & Faloutsos, M. (2004).** Long-range dependence: ten years of Internet traffic modeling. *IEEE Internet Computing*, 8(5), 57-64.
- 10 **Krunz, M. & Matta, I. (2002).** Analytical investigation of the bias effect in variance-type estimators for inference of long-range dependence. *Computer Networks*, 40(3), 445-458.
- 11 **Ledesma, S. & Liu, D. (2000).** Synthesis of fractional Gaussian noise using linear approximation for generating self-similar network traffic. *ACM Computer Communication Review*, 30(2), 4-17.
- 12 **Ledesma, S., Liu, D. & Hernandez, D. (2007).** Two approximation methods to synthesize the power spectrum of fractional Gaussian noise. *Computational Statistics and Data Analysis*, 52(2), 1047-1062.
- 13 **Leland, W. E. & Wilson, D. V. (1991).** High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. *IEEE Conference on Computer Communications INFOCOM*, Florida, USA, 1360-1366.
- 14 **Leland, W. E., Taqqu, M. S., Willinger W. & Wilson, D. V. (1994).** On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1), 1-15.
- 15 **Mandelbrot, B. B. & Wallis, J. R. (1969).** Some long-run properties of geophysical records. *Fractals in the earth sciences*, New York: Plenum Press.

- 16 **Michiel, H. & Laevens, K. (1997).** Teletraffic engineering in a broad-band era. *Proceedings of the IEEE*, 85(12), 2007-2033.
- 17 **Oppenheim, A. V. & Schaffer, R. W. (1989).** *Discrete-time signal processing*, Upper Saddle River, N.J., Prentice Hall.
- 18 **Paxson, V. (1997).** Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. *ACM SIGCOMM Computer Communications Review*, 27(5), 5-18.
- 19 **Paxson, V. & Floyd, S. (1995).** Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3), 226-244.
- 20 **Purczynski, J. & Wlodarski, P. (2005).** On fast generation of fractional Gaussian noise. *Computational Statistics and Data Analysis*, 50(10), 2537-2551.
- 21 **Ramirez, J. C. & Torres, R. D. (2006).** Local and Cumulative Analysis of Self-similar Traffic Traces, *IEEE Proceedings of the 16th International Conference on Electronics, Communications and Computers (CONIELECOMP'06)*, 37(1), 27-33.
- 22 **Roughan, M., Veitch, D. & Abry, P. (1998).** On-line estimation of the parameters of long-range dependence. *Proceedings of GLOBECOM'98*, Sydney, Australia, 3716-3721.
- 23 **Stoev, S., Taqqu, M. S., Park, C., Michailidis, G. & Marron, J. S. (2006).** LASS: a tool for the local analysis of self-similarity. *Computational Statistics & Data Analysis*, 50(9), 2447-2471.
- 24 **Tsybakov, B. & Georganas, N. D. (1998).** Self-similar processes in communications networks. *IEEE Transactions on Information Theory*, 44(5), 1713-1725.
- 25 **Veitch, D. & Abry, P. (1998).** Wavelet analysis of long-range dependent traffic. *IEEE Transactions on Information Theory*, 44(1), 2-15.
- 26 **Veitch, D. & Abry, P. (1999).** A wavelet based joint estimator of the parameters of long-range dependence. *IEEE Transactions on Information Theory*, 45(3), 878-897.
- 27 **Zukerman, M., Neame T. D. & Addie, R. G. (2003).** Internet traffic modeling and future technology implications. *IEEE Conference on Computer Communications INFOCOM*, California, USA, 587-596.



Sergio Ledesma received his Ph.D. in 2001 from Stevens Institute of Technology from the Department of Electrical and Computer Engineering. He currently holds a professor and researcher position at the School of Engineering from the University of Guanajuato. He is the author of the software *Neural Lab* to simulate artificial neural networks and *Wintempla* to develop Win32 applications. He worked as software engineering in the United States for Barclays Global Investors and other fines corporations.



Gustavo Cerda Villafaña received his B.S. from the Mechanical Engineering Department of the University of Guanajuato in 1994. He got his M.S. from the Electrical Engineering Department from the same university in 2000. He received his Ph.D. from the University of Bristol, England in 2005. He is currently working on the development of systems in artificial intelligence and statistical signal processing.



J. Gabriel Avina-Cervantes received the Engineering degree in Electronics and Communications from the University of Guanajuato in 1998, the M.S. degree in Electrical Engineering from University of Guanajuato in 1999, the Ph.D. in Informatics and Telecommunications from the “Institut National Polytechnique de Toulouse” and the LAAS-CNRS, France, in 2005. His research interests include artificial vision for outdoor robotics, pattern recognition, object classification and image processing. He is currently a researcher and a full time professor at the University of Guanajuato, at the Engineering Division of the Campus Irapuato-Salamanca.



Donato Hernández Fusilier received his M. S. (1990) in Electrical Engineering from the school of engineering (FIMEE) of University of Guanajuato (Mexico), and his B. S. (1986) in Communications and Electronics Engineering from the same university. Currently, he holds an associated professor position at the DICIS of the University of Guanajuato. His major research interests are in Neural Networks and signals optics.



Miguel Torres Cisneros. Received the Engineering degree in Electronics from the University of Guanajuato in 1987, the M.Sc. degree from the “Centro de Investigaciones en Óptica (CIO)” in 1991, and his D.Sc. from the “Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)” in 1997 with Honors. He was Visitor Researcher at the Universities of Dayton (UD) and of Central Florida (UCF) in 2002 and 2009, respectively. He has worked in several Mexican Universities as the “Tecnológico de Monterrey” and “the Universidad de las Américas”, and in the last 14 years in the Universidad de Guanajuato as a full professor. He has 34 journal papers, 1 patent, 93 proceedings and 7 divulgation papers. Belongs to the National Research System (SNI) from 1992, currently level 2, to PROMEP from 1999 and to the National Academy of Sciences from 2006. His research interests include Non Linear Optics, Image Processing, Materials synthesis and characterization and micro fabrication.