Pitfalls in Long Range Dependence Testing and Estimation

Sándor Molnár, Trang Dinh Dang
High Speed Networks Laboratory, Department of Telecommunications and Telematics
Budapest University of Technology and Economics, H–1117, Pázmány Péter sétány 1/D, Budapest, Hungary
Tel: (361) 463 3889, Fax: (361) 463 3107, E-mail: {molnar, trang}@ttt-atm.ttt.bme.hu

***Abstract*-Measured traffic traces from live packet networks often contain non-stationary effects like level shifts or polynomial trends. In these cases the popular tests for long-range dependence (LRD) can result in wrong conclusions and unreliable estimation of the Hurst parameter. In this paper we investigate the implications of these effects on several LRD tests. The use of these results can be utilized to avoid pitfalls in LRD traffic modeling. Our results are supported by both analytical and simulation studies with examples taken from traffic analysis of a live ATM network.**

I. Introduction

A very promising approach to capture the bursty nature of packet traffic in a parsimonious manner is to use *fractal traffic models* [13], [19]. The dynamics of these models are governed by *power-law* distribution functions and *hyperbolically* decaying autocorrelation [19]. The important characteristics of these models are *self-similarity* and *long-range dependence* [8], [13].

Self-similar stochastic processes have been defined in a number of ways in the literature [8], [13], [19]. In practice the most important class of these processes is that of long-range dependent (LRD) processes [8], [13]. LRD has been detected as a widespread property of packet network traffic, e.g. Internet traffic [12], [18]. In this paper we consider this class of self-similar processes defined in the next section.

From a practical point of view the important issues are the identification of LRD phenomena and the estimation of LRD parameters, especially the estimation of the Hurst parameter. Unfortunately, testing for LRD of measured data is not possible by simply checking the definitions. Instead, we can use some methods for testing the presence of some characteristics of the data which can or cannot support LRD, and also can or cannot give a reliable estimate of the Hurst parameter. Moreover, if all methods support the assumption of the presence of LRD with some H parameter it is still possible that this observation is caused by non-stationarities present in the data and are not due to the LRD. In this case it is possible to end up with wrong conclusions and build wrong models. In order to avoid such pitfalls we address this problem in this paper and give analytical and simulation investigations of the effects of different non-stationarity phenomena in the data.

The issue is not new and also addressed in the hydrology literature (e.g. [9]) after the application of LRD processes in the modeling of natural storage systems by Hurst [7], Mandelbrot and others [11]. However, after the invent and first application of LRD processes in the teletraffic research a number of papers have been published just by blind application of some LRD tests assuming the stationarity for hours of the traffic and taking no care of this important question.

The research was supported by the Inter-University Centre for Telecommunications and Informatics (ETIK).

We note that the problem was also addressed in the recent teletraffic literature, e.g. in [2], [4], [6], [15], [13] and also see the related references in [19] but stationarity tests and the validation techniques of fractal models have not been widely applied in today's teletraffic practice.

There are some ways to deal with this problem. One practical solution is based on the notion of *local stationarity*. Here we assume stationarity only over a short period of time and therefore our model parameters are valid only for this period and should be updated in the next period. A measurement-based approach with periodic real-time parameter estimation is a possible solution. Local stationarity with traditional models can also be used to capture the observed characteristics [17].

An alternative but rather difficult approach is to use *non-stationary models*, e.g. [5].

Some authors argue that this topic is somewhat philosophical from an application point of view [6], [9]. Indeed, if the modeling alternative can provide useful practical tools to dimension our networks then this can be a non-questionable proof for a proposed model. However, if more alternatives can work then we may prefer the parsimonious one which is a nice feature of fractal models. We believe that besides these factors the final choice of the proposed model and understanding about the nature of network traffic should be made not only by the analysis of the measured data but our *a priori* knowledge about the traffic generation process.

The contribution of this paper is to reveal the implications of the most important non-stationary effects which occur in practice on the most frequently used LRD tests in order to have a good understanding of these phenomena and to investigate the robustness of these tests against non-stationarity effects. The practical use of our findings is to support teletraffic engineers with guidelines to the effect that actual non-stationarities are not mistaken for stationary fractal behaviour.

Section II briefly introduces LRD and the methods of tests and estimations. Our analytical investigations for the tests of variance-time plot and R/S plot with level shifts and linear trends are given in Section III. Our simulation study with several examples is presented in Section IV, and Section V concludes the paper with some useful guidelines for LRD testing.

II. Practical Challenges in LRD Testing

This section gives a short overview of LRD processes and introduces the most frequently used test methods which are analyzed in the paper.

Let $X = (X_k : k \geq 0)$ be a covariance-stationarity process with autocorrelation function $r(k)$. X is said to exhibit *long-range dependence (LRD)* [2], [8] if $r(k) = k^{2H-2}L(k)$ as

$k \to \infty$, $0.5 < H < 1$, where $L(k)$ is slowly varying at infinity, i.e., $\lim_{k\to\infty}[L(tk)/L(k)] = 1,\ t > 0$.

As discussed in the previous section the tasks for testing of LRD and the estimation of Hurst parameter are not simple in practice. The main problem is that it is rather difficult to distinguish between non-stationary processes and stationary LRD processes due to the fact that LRD processes appear to have local trends, cycles, etc., many of the characteristics of non-stationary processes. These properties disappear after some time but if we have a finite and sometimes also short data set this identification is almost impossible. With a longer data set this identification becomes easier but we know for sure that in a long measured data non-stationary effects are present due to the daily cycles of traffic characteristics. The assumption about stationarity with high reliability may only be supported in the *busy periods* of the traffic.

There are methods developed to test for stationarity (e.g. [2], [14], [17]) and to distinguish between LRD and non-stationarity effects (e.g. [2], [10], [16]) but the application of these tests is not easy in practice. Moreover, such tests can seldom support their results with high reliability. In the next section the impacts of some kinds of non-stationarity effects on some LRD tests are analytically investigated. We are concerned with four widely used tests: the variance-time plot, the R/S analysis, the periodogram plot and the wavelet based H-estimator. Detailed descriptions of these methods can be found e.g. in [1] and [2].

III. Analytical Investigations

In this section our analytical study which shows how some non-stationary effects can change the results of some widely used LRD tests is briefly presented. Three cases are examined: variance-time plot of LRD data with level shift, with linear trend, and R/S analysis of LRD data with level shift.

Consider an $\{X_1, X_2, ...X_n\}$ series which is LRD with Hurst parameter H. To make the later calculation simpler two assumptions are made: (1) n is large enough so that aggregated series of $\{X\}$ used in computation of variance-time plot still contains large amount of data; (2) the mean of $\{X_i, 1 \le i \le n\}$ is zero. The second assumption can be taken into account because non-zero mean of LRD data does not change the result of LRD tests.

A. Variance-time plot of LRD data with level shift

The *variance-time plot* is the log-log plot of the variance of data versus the aggregation level [2]. In the case of LRD processes it can be proven [2] that

$$\text{Var}(X^{(m)}) = m^{2H-2}\ \text{Var}(X) \qquad \text{as } m \to \infty, \tag{1}$$

where m denotes the aggregation level. Therefore the variance-time plot of a LRD process with Hurst parameter H should be a straight line with slope $(2H-2)$ at large values of the aggregation level m.

By adding a level shift to series X we get the new series denoted by X^{LS}. The level shift is assumed to have a simple shape: it has two states of value 0 and K_{LS} separated at the center of the investigated data. We have proven [3] that in this case

$$\text{Var}\left(X^{LS(m)}\right) \approx \frac{\text{Var}\left(X^{LS}\right) - K_{LS}^2/4}{m^{2-2H}} + \frac{K_{LS}^2}{4}. \tag{2}$$

Plotting $\log\left[\text{Var}\left(X^{LS\ (m)}\right)\right]$ against $\log m$ we get a convex curve bounded by two lines. The line with slope $2H-2$ and ordinate $\log\left[\text{Var}\left(X^{LS}\right) - K_{LS}^2/4\right]$ as $m \to 0$ and a horizontal line with ordinate $K_{LS}^2/4$ as $m \to \infty$. This shows that the variance-time plot of LRD data with level shift has a convex curve and asymptotically approaches a horizontal line. The estimation of H for LRD processes should be performed at large m (in theory as $m \to \infty$). Therefore we can conclude that the estimation is highly destroyed in the presence of level shifts. Further details about this distortion are demonstrated by examples given in Section IV.

B. Variance-time plot of LRD data with linear trend

A linear trend was added to the LRD data X with the maximum value denoted by K_L. This new data series is denoted by X^L. We have proven [3] that in the case of LRD process with this linear trend we have

$$\text{Var}\left(X^{L(m)}\right) \approx \frac{\text{Var}(X^L) - C_1}{m^{f_L(m)}}, \tag{3}$$

where the constant C_1 is independent of m and $f_L(m)$ is a computable function of m [3]. Equation 3 shows that the presence of a linear trend in LRD data turns the result of variance-time plot to be quite different from its original form. Plotting $\log\left[\text{Var}(X^{L\ (m)})\right]$ versus $\log m$ instead of a straight line with slope $(2H-2)$ we can observe a curve described by $f_L(m)$. The estimation of the Hurst parameter of LRD from the variance-time plot should be done by fitting a regression line to the plot at large values of m, so from 3 as m tends to infinity we get [3]

$$\text{Var}\left(X^{L(m)}\right) \approx \frac{\text{Var}(X^L) - C_1}{m^{2-2H}} + C_2 K_L + \frac{7\,K_L^2}{12} \tag{4}$$

Equation 4 concludes that the variance-time plot of a LRD process with linear trend asymptotically approaches a horizontal line with ordinate $C_2 K_L + 7\,K_L^2/12$, where the constant C_2 is independent of m. The variance-time plots of the LRD process and a process with no LRD are no longer distinguishable in the presence of a linear trend. For more details see our examples in Section IV.

C. R/S plot of LRD data with level shift

The R/S analysis of an $\{X_1, X_2, \ldots, X_n\}$ data series is defined by the log-log plot of the *rescaled adjusted range* (R/S ratio) versus the actual data window size d [2]. For a certain window size d we consider data $X_i,\ off < i \le d$. The R/S value is given by:

$$\frac{R}{S} = \frac{\max\{(W_i - W_j);\ i,j = 1,2,\ldots,d\}}{\sqrt{\text{Var}(X_{off,d})}}, \tag{5}$$

where $X_{off,d}$ denotes the considered sub-series $\{X_{off+1}, X_{off+2}, \ldots, X_{off+d}\}$ and $W_i = \sum_{k=1}^{i}(X_{off+k} - \bar{X}_{off,d})$ where $\bar{X}_{off,d}$ denotes the mean of $X_{off,d}$. With a value of d we calculate several R/S ratios by sliding the window of size d throughout the set of X series. The R/S ratio of LRD data has the following characteristics $R/S \sim C_H d^H$ as $n \to \infty$, where C_H is an finite positive constant independent of d [2].

According to the definition of the R/S ratio we observed that this ratio does not change if the data window with size d does

not cover the level shift. There is a different case when the data window contains the level shift. The simple case when the location of the shift is placed at the center of the window is concerned:

$$X^{*\ LS}_{off+k} = \begin{cases} X_{off+k} & \text{if } k \le \lfloor n/2 \rfloor \\ X_{off+k} + K_{LS} & \text{if } k > \lfloor n/2 \rfloor \end{cases},$$

where $k = 1, 2, \ldots, d$ and (*) means that it only relates to those d-windows mentioned above. The following holds [3] for large enough values of d:

$$\left(\frac{R}{S}\right)^{*LS} \approx \frac{d\ K_{LS}/4}{\sqrt{\mathrm{Var}(X) + K^2_{LS}/4}} = d\ C_3, \qquad (6)$$

where C_3 is a constant independent of d.

These points create a separate part on the log-log plot which should be placed closely around a straight line with slope 1. The other large cluster of points remains at the same place as before adding level shift and this part of the R/S plot of LRD data with level shift looks similarly like the R/S plot of the original LRD data.

This result shows that the R/S plot can also be used for detection of level shifts in the data. Moreover, the linear part with slope 1 in the plot should be disregarded in the estimation of Hurst parameter of LRD processes. In this way in the cases when this separation is feasible we can make a reliable estimate of H even in the presence of level shifts.

IV. Simulations and Analysis of Measured ATM Traffic

A. Trend types

The analysis of measured packet traffic can reveal various deterministic changes in the data on different time scales. These traffic variations are not stochastic by nature but rather caused by deterministic mechanism like protocols. These mechanisms can, for example, introduce quasi-periodic patterns in the traffic data which can be, if not detected and removed, the cause of several statistical pitfalls, e.g. the conclusion of slowly decaying correlations.

On higher time scales a regular character of the traffic due to daily or weekly variations can be observed. These traffic trends should also be identified and removed prior to any statistical analysis. These are not easy but important parts of a comprehensive statistical analysis. An alternative approach is to use tests which are robust against these non-stationary effects.

Different trend models [3] are candidates for investigations, e.g. linear trend, parabolic trend, exponential trend, logistical trend or Gompertz trend, etc. We have chosen the non-stationary effects and trends which are frequently observed in practice. These are the *level shift*, which can be observed when during our traffic measurements suddenly a new source starts to emit traffic to the aggregation and the *linear and parabolic trends*, which can be observed in daily traffic variations.

A sample series (containing 32,768 samples) of Fractional Gaussian Noise (FGN) [2] was used as a reference data set exhibiting LRD. The Hurst parameter was set to 0.7 and the series is denoted by 0.7-FGN. In the next step linear, parabolic trends and level shift were added to this data set denoted by 0.7-FGN_L, 0.7-FGN_P, and 0.7-FGN_LS, respectively. Table I gives more information about these data sets.

TABLE I

Characteristics of investigated data sets ($\hat{\mu}$ and $\hat{\sigma}^2$ denote the sample mean and the sample variance, respectively).

Data sets	$K_{L,LS,P}$	$\hat{\mu}$	$\hat{\sigma}^2$
0.7-FGN	-	10	10
0.7-FGN_L	5	12.5	17.78
0.7-FGN_LS	5	12.5	18.22
0.7-FGN_P	5	11.66	34.53

A series of ATM cell arrivals obtained from a real-time traffic measurement on the Swedish University NETwork (SUNET) [13] was also analyzed. Data traces were collected in 1996 based on a custom-built measurement tool which is able to record more than 8 millions consecutive cell arrivals. In our tests the traces of the number of cell arrivals in a 1ms time window were considered. The analysis of these data traces can illustrate the non-stationary effects in LRD estimation of real traffic.

B. Empirical results

Variance-time plot The variance-time plots of the FGN series are presented in Fig. 1. The test result of 0.7-FGN produces a straight line, which yields the estimated Hurst parameter value 0.7. That is the exact value we expected. Fig. 1 also shows variance-time plots of the 0.7-FGN set with different non-stationary effects. As we can observe all curves are convex which gives no information about the LRD property of the original FGN data. If one tries to fit a line for the large values of m, as it is usually done in such an analysis, the result would be misleading due to the distorted slope of the curve. This result also justifies our analysis conclusions in Subsection III-A and III-B, and illustrates that both 07-FGN_LS and 07-FGN_L curves seem to converge to a horizontal line.

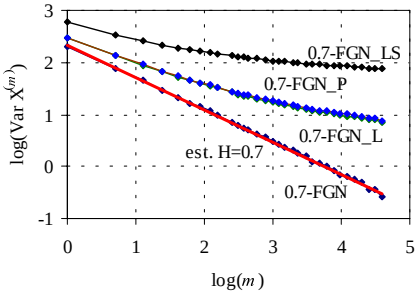

Fig. 1. The variance-time plot of 0.7-FGN data series (the curves of 0.7-FGN_L and 0.7-FGN_P nearly coincide)

The result of variance-time analysis of the SUNET ATM data is presented in Fig. 2. The measured ATM traffic is bursty in nature and although several pre-processing procedures were done in this trace it is difficult to detect a certain trend. However, the curve is very similar to those obtained with level shift or trends in Fig. 1. The estimation of H applied in such a variance-time plot can produce misleading results.

R/S plot Fig. 3 and 4 show the R/S plots of the 0.7-FGN and the 0.7-FGN_LS data sets, respectively. In the case of the 0.7-

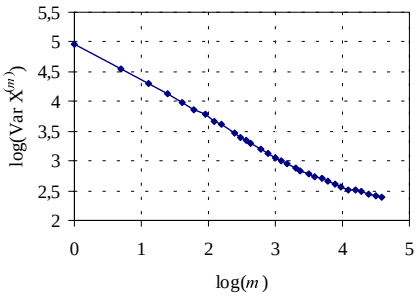

Fig. 2. The variance-time plot of the SUNET ATM data

FGN data set the estimation of the Hurst parameter returns the exact value of H set to this series. However, the interesting

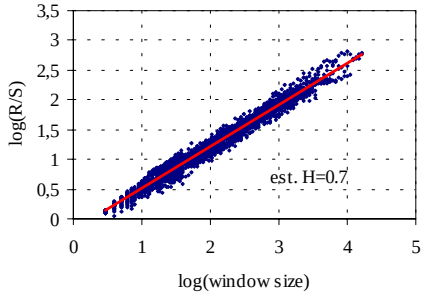

Fig. 3. The R/S analysis of the 0.7-FGN data set

result is found in the plot of FGN series with level shift 0.7-FGN_LS. On one hand, the plot seems to be constructed from two parts which are independent of each other. The lower part looks exactly like the R/S plot of the original set as in Fig. 3. On the other hand, the upper part is similar to a line of slope 1. This result is in good agreement with our analytical results presented in Subsection III-C.

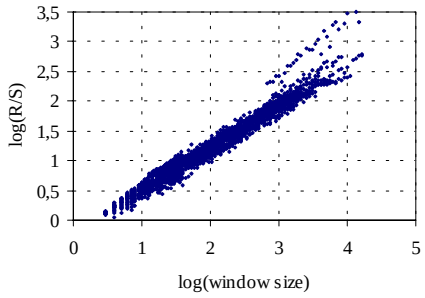

Fig. 4. The R/S analysis of the 0.7-FGN_LS data set

An illustrative example of such an effect from practice can be seen in the R/S plot of the SUNET ATM series (Fig. 5). The plot contains a break point where the slope of the curve changes approximately placed in the middle of the figure. If one tries to estimate H from the upper part of the plot it will result in a wrong value as we demonstrate it in the following. Fig. 6 shows the R/S plot of a subset of the SUNET ATM set. The subset is gained from the original set after erasing some suspected non-stationary parts of the data set. In Fig. 6 the part of the curve with the higher slope disappeared and the lower part continues growing nearly as a straight line. An explanation of

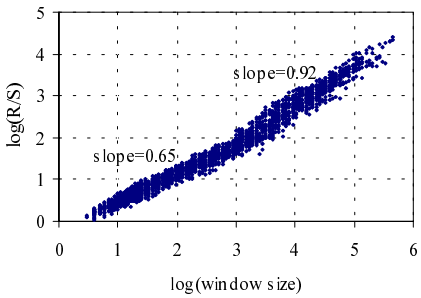

Fig. 5. The R/S plot of the SUNET ATM series

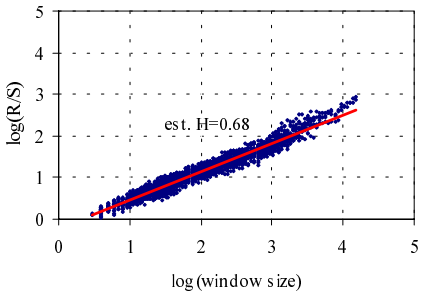

Fig. 6. The R/S plot of the "stationary" subset of the SUNET ATM data set

this phenomenon is the possible presence of several local level shifts in the original SUNET ATM data. The result also demonstrates that the important part for LRD parameter estimation is distorted by level shifts.

Periodogram plot In the frequency domain adding deterministic trend to a signal produces the increase of low frequency components. Thus we were not surprised when observing the rise of the lower tail of the periodograms under the influence of different trends. As an example the periodogram of 0.7-FGN data without and with linear trend are presented in Fig. 7 and Fig. 8, respectively. Since periodograms at low frequencies should be

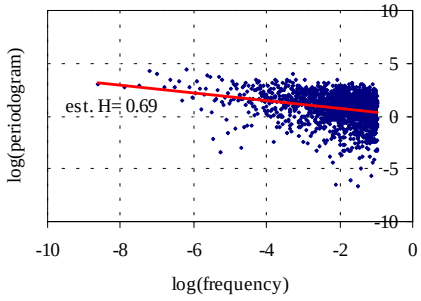

Fig. 7. The periodogram of the 0.7-FGN set

counted for estimation of the Hurst parameter, the presence of trends in LRD data destroys the testing and estimating capability of the periodogram plot.

Wavelet-based estimator The LRD test based on the wavelet transformation, called the logscale diagram [1], is investigated. Fig. 9 presents the result of the 0.7-FGN set provided by the Logscale diagram. The estimate of H is 0.71 with confidence interval (0.7, 0.72) which is close to the exact value of H.

As proven in [1] the influence of polynomial trends on this

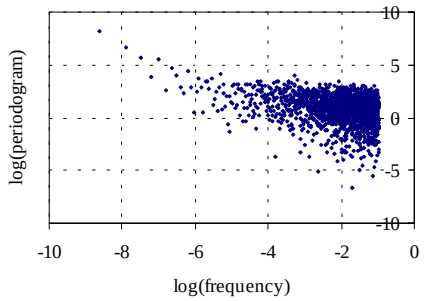

Fig. 8. The periodogram of the 0.7-FGN_L data set

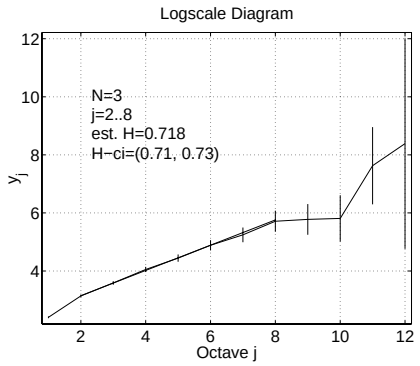

Fig. 9. Logscale diagram of the 0.7-FGN data set

kind of LRD test can be avoided by an adequate choice of the vanishing moments of the wavelet function. Our empirical work has justified this observation. Moreover, our simulation also shows that the logscale diagram is still robust in the presence of level shifts. As seen in Fig. 10 the level shift added to the 0.7-FGN set slightly changes the result: the estimation of H is 0.72 with confidence interval (0.713, 0.729).

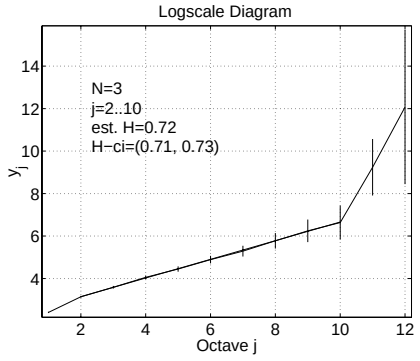

Fig. 10. Logscale diagram of the 0.7-FGN_LS data set

V. Conclusions

Our analytical and simulation analysis has shown that the presence of different non-stationary effects (level shifts, linear and polynomial trends) in the data can deceive several LRD tests. The results were also demonstrated by practical examples from the analysis of measured ATM traces.

In the case of the variance-time plot and the periodogram these effects result in a poor estimate of the Hurst parameter. Moreover, the estimated results can be confused with results of processes having short-range dependence with non-stationary effects. We conclude that the variance-time plot and the periodogram methods should not be used without a stationarity and trend analysis.

The R/S analysis can reveal the presence of the level shifts, therefore it is a good candidate method for a test. However, the Hurst parameter estimation of R/S test without the removal of points caused by the level shift should also be avoided.

The wavelet-based method provides a very robust estimation of H even in the presence of level shifts or trends. We recommend the Logscale diagram for the estimation of the Hurst parameter of LRD processes in the possible presence of the investigated non-stationary effects.

References

[1] P. Abry and D. Veitch, "Wavelet Analysis of Long-Range Dependent Traffic," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.

[2] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, One Penn Plaza, New York, NY 10119, 1995.

[3] T.D. Dang and S. Molnár, "On the effects of non-stationarity in long-range dependent tests," Tech. Rep., Budapest University of Technology and Economics, 1999.

[4] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russel, and F. Toomey, "Statistical issues raised by the Bellcore data," in *11th Teletraffic Symposium*, Cambridge, 23-25 March 1994.

[5] N. G. Duffield, W. A. Massey, and W. Whitt, "A nonstationary offered-load model for packet networks," in *Sel. Proc. of the 4th INFORMS Telecomm. Conf.*, 1999.

[6] A. Erramilli, P. Pruthi, and W. Willinger, "Self-similarity in high speed network traffic measurements: Fact or artifact?," in *Proc. of the 12th Nordic Teletraffic Seminar, NTS12*, Espoo, Finland, 22-24 Augustus 1995.

[7] H. E. Hurst, "Long-term storage capacity of reservoirs," *Proc. Amer. Soc. Civil Eng.*, vol. 76(11), 1950.

[8] D.L. Jagerman, B. Melamed, and W. Willinger, "Stochastic Modeling of Traffic Processes," in *Frontiers in Queueing*, pp. 271–370. CRC Press, 1997.

[9] V. Klemes, "The Hurst phenomenon: A Puzzle?," *Water Resources Research 10*, pp. 675–688, 1974.

[10] H. Kunsch, "Discrimination between monotonic trends and long-range dependence," *J. Appl. Prob.*, vol. 23, 1986.

[11] B.B. Mandelbrot and J.W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Rev.*, vol. 10, pp. 422–437, 1968.

[12] S. Molnár and T.D. Dang, "Scaling Analysis of IP Traffic Components," in *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, USA, 18-20 September 2000.

[13] S. Molnár and I. Maricza eds., "Source Characterization in Broadband Networks," Interim Report, COST 257, Vilamoura, Portugal, January 1999.

[14] S. Molnár and A. Gefferth, "On the Scaling and Burst Structure of Data Traffic," in *8th Int. Conference on Telecommunication Systems, Modelling and Analysis*, Nashville, Tennessee, USA, 9-12 March 2000.

[15] S. Molnár, A. Vidács, and A.A. Nilsson, "Bottlenecks on the Way Towards Fractal Characterization of Network Traffic: Estimation and Interpretation of the Hurst Parameter," in *Proc., PMCCN'97*, Tsukuba, Japan, 1997, pp. 125–144.

[16] V. Teverovski and M. Taqqu, "Testing for long-range dependence in the presence of shifting means or a slowly declining trends, using variance type estimator," *J. of Time Series analysis*, vol. 18, no. 3, pp. 279–304, 1997.

[17] S. Vaton and E. Moulines, "A Locally Stationary Semi-Markovian Representation for Ethernet LAN Traffic Data," in *IFIP TC6/WG6.2 4th Int. Conference on Broadband Communications*, Stuttgart, Germany, 1-3 April 1998.

[18] A. Veres, Zs. Kenesi, S. Molnár, and G. Vattay, "On the Propagation of Long-Range Dependence in the Internet," in *ACM SIGCOMM 2000*, Stockholm, Sweden, 28 August - 1 September 2000.

[19] W. Willinger, M.S. Taqqu, and A. Erramilli, "A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High-Speed Networks Stochastic Networks:Theory and Applications," in *Royal Statistical Society Lecture Notes Series.* 1996, vol. 4, Oxford University Press.