

LASS: a tool for the local analysis of self-similarity

Stilian Stoev
Boston University

Murad S. Taqqu
Boston University

Cheolwoo Park
SAMSI

George Michailidis
University of Michigan, Ann–Arbor

J. S. Marron
University of North Carolina, Chapel Hill

January 30, 2004

Abstract

The Hurst parameter H characterizes the degree of long-range dependence (and asymptotic self-similarity) in stationary time series. Many methods have been developed for the estimation of H from data. In practice, however, the classical estimation techniques can be severely affected by non-stationary artifacts in the time series. In fact, the assumption that the data can be modeled by a stationary process with a single Hurst exponent H may be unrealistic.

We focus on practical issues associated with the detection of long-range dependence in Internet traffic data and develop two tools designed to address some of these issues. The first is an animation tool which is used to visualize the local dependence structure. The second is a statistical tool for the *local analysis of self-similarity* (LASS). The LASS tool is designed to handle time series that have long-range dependence and are long enough that some parts are essentially stationary, while others exhibit non-stationarity, which are either deterministic or stochastic in nature. The tool uses wavelets to analyze the local dependence structure in the data over a set of windows. It can be used to visualize local deviations from self-similar, long-range dependence scaling and to provide reliable *local* estimates of the Hurst exponents. The tool, which is illustrated by using a trace of Internet traffic measurements, can also be applied to economic time series.

We also develop a median-based wavelet spectrum which can be used to obtain robust local or global estimates of the the Hurst parameter that are less susceptible to local non-stationarity. We make the software tools freely available and describe their use in an appendix.

1 Introduction

In the past decade there has been a growing interest in modeling the traffic in modern computer telecommunication networks. A number of studies (Leland, Taqqu, Willinger and Wilson (1993), Paxson and Floyd (1995) and the collection of papers in Park and Willinger (2000)) have shown that the classical telephony models based on the homogeneous Poisson process apply to neither the Ethernet traffic nor the traffic in wide area networks such as the Internet.

As shown by Leland *et al.* (1993), among others, network traffic exhibits self-similarity and asymptotic self-similarity, consistent with long-range dependence. A stochastic process

$X = \{X(t)\}_{t \in \mathbb{R}}$ with zero mean is said to be self-similar with self-similarity parameter $H > 0$, if for all $c > 0$,

$$\{X(ct)\}_{t \in \mathbb{R}} =_d \{c^H X(t)\}_{t \in \mathbb{R}}, \quad (1.1)$$

where $=_d$ means equal *marginal* and *finite-dimensional distributions*. Relation (1.1) implies that the process $\{X(t)\}_{t \in \mathbb{R}}$ appears, up to a multiplicative constant, to be statistically the same at all time scales.

The fractional Brownian motion (FBM) process $B_H = \{B_H(t)\}_{t \in \mathbb{R}}$ is a self-similar Gaussian process with stationary increments and self-similarity parameter $H \in (0, 1)$ (see Taqqu (2003), Beran (1994), Mandelbrot and Van Ness (1968)). FBM has been one of the most successful macroscopic models for fluctuations in the traffic arrival process. That is, at moderately large time scales (about 1 sec) the aggregate network traffic can be often well-approximated by a Gaussian process with stationary increments. The only self-similar Gaussian process with stationary increments is the fractional Brownian motion. Therefore, due to the self-similarity property of network traffic the FBM process is a natural candidate for a traffic model. Furthermore, the FBM can be also interpreted as a physical model since it appears as the limit process in many structural traffic models such as the ON/OFF source model and the infinite source Poisson model (see, Taqqu, Willinger and Sherman (1997), Mikosch, Resnick, Rootzen and Stegeman (2002), Taqqu (2002) and the references therein).

Fractional Brownian motion is not an adequate model at all scales as shown, for example, by the recent extensive study of Hernández-Campos *et al.* (2004). It applies only to traffic fluctuations at intermediate time scales (e.g. from 1 second to 1 hour). At larger time scales non-stationary diurnal effects and deterministic trends in traffic start to play a dominant role. On the other hand, at very small time scales (e.g. below 1 second) the fluctuations in network traffic are no longer exactly statistically self-similar. They also possess non-Gaussian, skewed distributions and intricate dependence structure, which often depends on the type of the network (e.g. backbone core link, small ISP edge link, corporate link or university link), the dominant network protocols (e.g. TCP, UDP, NNTP, HTTP etc.), the dominant applications (e.g. web, mail, file-sharing, voice and video, etc.) and the user behavior.

Emerging applications and new protocol features can seriously affect the properties of network traffic and its evolution over time. The fluctuations of the Internet traffic cannot be viewed only as a purely physical phenomenon detached from technological and in fact social factors. This makes traffic modeling a very difficult and novel challenge, which is somewhat different from the challenges posed by natural phenomena in physics, hydrology or biology, for example.

The rapid technological development of the network hardware leads to greater link speeds. This in turn yields extremely large amounts of traffic data to analyze, much of which is non-stationary. Hence the need for better statistical tools.

In this work, we propose a statistical methodology for analyzing the local self-similar scaling in real data. The focus is on time series that have strong dependence, and are long enough that some parts are essentially stationary, while others exhibit non-stationarity, that are either deterministic or stochastic in nature. As shown in Stoev, Taqqu, Park and Marron (2004), some of the best available techniques to measure the Hurst parameter in data may be misled by non-stationary artifacts. Some of these non-stationarity effects can often be alleviated by looking at data locally in time, over a moving window. We present two wavelet-based tools for the local analysis of the Hurst parameter in Internet traffic. Our emphasis is on exploratory

analysis of the local scaling in data, in the spirit of the scale-space approach. That is, we focus on describing data and detecting statistically significant features locally in time and on a variety of scales, rather than on suggesting a rigid model. This type of analysis can be particularly useful in the context of networking, where often models fail to capture interesting details and nonetheless practitioners are faced with the challenge to understand the structure of the data. A theoretical treatment for the problem of testing stationarity of the local Hurst estimates can be found in Veitch and Abry (1999a).

The paper is organized as follows. In Section 2, we first briefly review the definition of long-range dependence and a basic model – the fractional Brownian motion. Then, we present the wavelet spectrum and sketch its use for the estimation of the Hurst long-range dependence parameter. In Section 2.3, we briefly discuss some practical limitations in the estimation of the Hurst parameter in long datasets, which possess non-stationarity. The tools presented in the rest of the paper address some of these limitations.

In Section 3, we describe an animation tool designed to visualize the local dependence structure of a time series and in Section 4, we present the Local Analysis of Self-Similarity (LASS) tool. The LASS tool provides insight into potential non-stationarity of a time series, through a combination of views. These views are motivated, using Internet traffic data. They include the global wavelet spectrum, coupled with a sequence of local wavelet spectra and local Hurst parameter estimates. Further effectiveness of the LASS combined views is demonstrated in Section 5, via a simulated example. Concluding remarks can be found in Section 6. The software tools are implemented in MATLAB. These tools can be used not only for Internet traffic data but also in the analysis of economic time series. They are freely available and their use is described in the appendix.

2 Global versus local analysis of self-similarity

We start by briefly reviewing some basic facts related to the long-range dependence phenomenon and fractional Brownian motion. Then, in Section 2.2, we present the wavelet spectrum of a stationary time series and discuss its use for the estimation of the Hurst parameter. In Section 2.3, we conclude by discussing some practical difficulties in the estimation of the Hurst parameter for long Internet traffic data sets. Some of these limitations can be bypassed by performing a *local* rather than *global* analysis of the long-range dependent scaling in data.

2.1 Long-range dependence: basic notions

We start by recalling some basic facts about long-range dependence. For more details, see Beran (1994) and Taqqu (2003).

Consider a second order stationary time series $Y = \{Y(k)\}_{k \in \mathbb{Z}}$ with mean zero. The time series Y is said to be *long-range dependent* if its auto-covariance function $r_Y(k) = \text{Cov}(Y(k), Y(0)) = \mathbb{E}Y(k)Y(0)$ decays slowly as a function of the lag k , so that the series $\sum_k r_Y(k)$ is not summable. Typically the long-range dependence is modeled by supposing a power-like decay of the covariances:

$$r_Y(k) = \mathbb{E}Y(k)Y(0) \sim c_Y |k|^{-\gamma}, \quad \text{as } k \rightarrow \infty, \quad 0 < \gamma < 1, \quad (2.1)$$

where $c_Y > 0$ and where \sim means asymptotic equivalence. Observe, that since $0 < \gamma < 1$, Relation (2.1) implies $\sum_k r_Y(k) = \infty$.

One can also model long-range dependence by imposing conditions on the spectral density f_Y of Y around the origin. Namely:

$$f_Y(\xi) \sim c_f |\xi|^{-\alpha}, \quad \text{as } \xi \rightarrow 0, \quad 0 < \alpha < 1 \quad (2.2)$$

where $c_f > 0$ and where $f_Y(\xi) := (2\pi)^{-1/2} \sum_{k \in \mathbb{Z}} e^{i\xi k} r_Y(k)$.

Under some smoothness assumptions, the conditions in (2.1) and (2.2) can be related and one has

$$\alpha = 1 - \gamma$$

(for more details, see, for example, Taqqu (2003) and the references therein). Furthermore, long-range dependent time series Y can be *asymptotically self-similar*. For example, if Y is Gaussian and satisfies Relation (2.1), one has that, as $n \rightarrow \infty$,

$$\left\{ \frac{1}{n^H} \sum_{k=1}^{[nt]} Y(k) \right\}_{t \in [0,1]} \xrightarrow{f.d.d.} \{B_H(t)\}_{t \in [0,1]}, \quad (2.3)$$

where $\xrightarrow{f.d.d.}$ means convergence in the sense of finite-dimensional distributions, $1/2 < H < 1$ and where B_H is the fractional Brownian motion (FBM) process (see, e.g. Theorem 7.2.11 in Samorodnitsky and Taqqu (1994), and Taqqu (1975)). The parameter H is called the Hurst parameter of the time series Y and it relates to the parameters γ and α in (2.1) and (2.2) as follows:

$$\gamma = 2(1 - H) \quad \text{and} \quad \alpha = 2H - 1. \quad (2.4)$$

The Hurst parameter quantifies the degree of long-range dependence as well as the asymptotic self-similarity scaling of the process Y .

The FBM is a self-similar Gaussian process with self-similarity parameter H and with stationary increments. In view of Relation (2.3), it plays a fundamental role in modeling long-range dependence. In practice, it is its increments that are used in modeling. The increments time series $G_H(k) := B_H(k) - B_H(k-1)$, $k \in \mathbb{Z}$ of the FBM process B_H are called *fractional Gaussian noise* (FGN) and have covariances:

$$\text{Cov}(G_H(k), G_H(0)) = \frac{\sigma^2}{2} \left(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H} \right),$$

where $\sigma > 0$. When $1/2 < H < 1$, the FGN G_H is long-range dependent and satisfies Relations (2.1) and (2.2) with parameters γ and α as in (2.4).

The Hurst parameter H is an important parameter, characterizing the long-term dependence structure of a time series. In the following section, we briefly present one of the most successful techniques to estimate H in practice.

2.2 Wavelet spectrum: estimation of the Hurst parameter

Wavelets have become a popular tool in the analysis of long-range dependence properties of network traffic. Here we shall briefly sketch the definition of the *wavelet spectrum* of a second

order stationary time series Y and its use for the estimation of the Hurst parameter H of Y . For more details, see Abry and Veitch (1998) and Abry, Flandrin, Taqqu and Veitch (2000). We also introduce here a *robust* version of the wavelet spectrum, which limits the effect of large fluctuations due to non-stationarity or heavy bursts in the traffic data.

Let $\psi(s)$ be a square integrable function. The function ψ is called an orthogonal *mother wavelet* if the collection of all its dyadic dilations and integer translates $\psi_{j,k}(s) := 2^{-j/2}\psi(2^{-j}s - k)$, $j, k \in \mathbb{Z}$ forms an orthonormal basis of $L^2(ds) = \{f(s), \int_{\mathbb{R}} f^2(s)ds < \infty\}$. It follows that any function $f(s) \in L^2(ds)$, admits the expansion

$$f(s) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} d_{j,k}(f) \psi_{j,k}(s), \quad \text{where } d_{j,k}(f) = \int_{\mathbb{R}} f(s) \psi_{j,k}(s) ds,$$

and where the last function series converges in the L^2 -sense. The coefficients $d_{j,k}(f)$ are called *detail* wavelet coefficients of the function f . The set $\{d_{j,k}(f)\}_{j,k \in \mathbb{Z}}$ is referred to as the *discrete wavelet transform* of the function f . The indexes j and k of the $d_{j,k}(f)$ s are called *scale* and *location*, respectively. The coefficient $d_{j,k}(f)$ captures the behavior of the signal f , localized at times about $2^j k$ and restricted to a frequency band about 2^j . Therefore wavelets present a rich time/frequency picture of the signal, which can be more informative than that of the classical Fourier analysis.

Orthogonal mother wavelets can be elegantly constructed using the framework of the *multiresolution analysis* of $L^2(ds)$ (for more details, see, for example, Ch. 6 in Daubechies (1992)).

To analyze a discrete time series $Y(k)$, $k = 1, \dots, N$ by using wavelets one typically uses Mallat's fast pyramidal algorithm. One obtains an array of wavelet coefficients $d_{j,k}$, $j = 1, \dots, J$, $k = 1, \dots, N_j$, where $J \approx \log_2(N)$ and where $N_j \approx N/2^j$. The $d_{j,k}$ s are the "detail" coefficients of a continuous time process \tilde{Y} , related to the time series Y (see e.g. Stoev, Taqqu, Park and Marron (2004) and Ch. 6 in Daubechies (1992), for more details).

The Hurst long-range dependence parameter of the time series $Y(k)$ appears naturally in the scaling of the *energy* of the wavelet coefficients $d_{j,k}$. In particular, for sufficiently large scales j , one has that

$$\log_2 \mathbb{E} d_{j,k}^2 \sim j(2H - 1) + C, \quad (2.5)$$

where C is a constant independent of k and where H denotes the Hurst parameter of the time series $Y(k)$.

The stationarity of the time series $Y(k)$ implies the stationarity in k of the $d_{j,k}$, for any fixed scale $j = 1, \dots, J$. Furthermore, although the $Y(k)$ s can be strongly dependent, the wavelet coefficients $d_{j,k}$ are essentially uncorrelated in k (see, e.g. Kim and Tewfik (1992) and Bardet, Lang, Moulines and Soulier (2000)). Therefore, one can estimate the left-hand side in (2.5) by using the statistics

$$S_j(Y) := \log_2 \left(\frac{1}{N_j} \sum_{k=1}^{N_j} d_{j,k}^2 \right) - g_{N_j}(j), \quad (2.6)$$

where $g_{N_j}(j)$ is a bias correction term of the order $(\ln(2)N_j)^{-1}$, as $N_j \rightarrow \infty$. This correction term compensates for the fact that the logarithm of the expectation is not equal to the expectation of the logarithm.

The set of statistics $S_j(Y)$, $j = 1, \dots, J$ is called the *wavelet spectrum* of the time series $Y(k)$, $k = 1, \dots, N$. At large scales j , the statistics S_j capture features in the low-frequency region of the Fourier spectral density. Therefore, in view of (2.2), the large scale part of the wavelet spectrum represents the long-range dependence behavior of the time series Y . On the other hand, the wavelet spectrum S_j at small scales j reflects the short-term dependence structure of Y .

The Hurst parameter H of Y can be estimated by using a linear regression of the wavelet spectrum S_j on the scales j over a range of sufficiently large scales going from j_1 to j_2 , where $1 \leq j_1 < j_2 \leq J$. Namely,

$$\hat{H}_{[j_1, j_2]} := \frac{1}{2} \sum_{j=j_1}^{j_2} w_j S_j(Y) + \frac{1}{2}, \quad (2.7)$$

where w_j s are such that $\sum_{j=j_1}^{j_2} w_j = 0$ and $\sum_{j=j_1}^{j_2} j w_j = 1$. In practice, the weights w_j , $j = j_1, \dots, j_2$ are carefully chosen in order to reduce the variance of the estimator \hat{H} (for more details, see Abry and Veitch (1998, 1999b)).

Alternatively, consider the statistics

$$\hat{H}_{[j, j+1]} := \frac{(S_{j+1}(Y) - S_j(Y))}{2} + \frac{1}{2}, \quad j = 1, 2, \dots, J-1, \quad (2.8)$$

where $S_{j+1}(Y) - S_j(Y)$ represents the local slope of the wavelet spectrum at scale j . In view of (2.5), for large j , we expect $S_j(Y) \approx j(2H - 1) + C$ and hence $\hat{H}_{[j, j+1]} \approx H$, provided that N_j is also sufficiently large. One can express the estimator $\hat{H}_{[j_1, j_2]}$ in (2.7) of the Hurst parameter in terms of the statistics $\hat{H}_{[j, j+1]}$:

$$\hat{H}_{[j_1, j_2]} = \sum_{j=j_1}^{j_2-1} v_j \hat{H}_{[j, j+1]} \quad (2.9)$$

by setting $v_j = w_{j+1} + \dots + w_{j_2}$, $j = j_1, \dots, j_2 - 1$. One has $\sum_{j=j_1}^{j_2-1} v_j = 1$.

Empirically, the statistics $\hat{H}_{[j, j+1]}$, $j = 1, \dots, J-1$ appear to be weakly correlated and, for large sample sizes, are well-approximated by a multivariate Gaussian distribution. Using this property, we propose, in Section 4.1 below, a statistical methodology which allows one to visualize how the choice of scales j_1 and j_2 affects the estimation of the Hurst parameter.

When the time series Y is Gaussian, so are the wavelet coefficients $d_{j,k}$. However, if Y is a heavy-tailed time series with infinite variance, then the $d_{j,k}$ will have infinite variance and consequently the statistics S_j may not be consistent. In such cases one can use alternative wavelet spectra such as:

$$S_j^\beta(Y) := \frac{2}{\beta} \log_2 \left(\frac{1}{N_j} \sum_{k=1}^{N_j} |d_{j,k}|^\beta \right) - g_{N_j}(j) \quad (2.10)$$

where $0 < \beta < 2$ is a free parameter or

$$S_j^{\log}(Y) := 2 \frac{1}{N_j} \sum_{k=1}^{N_j} \log_2 |d_{j,k}|. \quad (2.11)$$

Using these definitions of wavelet spectrum one can construct estimators \widehat{H}_β and \widehat{H}_{\log} , as in (2.7). Closely related estimators were shown to be consistent and asymptotically normal (see Pipiras, Taqqu and Abry (2001), Stoev, Pipiras and Taqqu (2002) and Stoev and Taqqu (2003)). In practice the estimators \widehat{H}_β and \widehat{H}_{\log} work rather well.

- *Robust, median-based wavelet spectrum*

One can also consider a wavelet spectrum focused on sample medians. This spectrum is robust with respect to large, rare fluctuations in the wavelet coefficients due to non-stationarity anomalies in the data. Its definition parallels that of the wavelet spectra in (2.6), (2.10) and (2.11). Let

$$S_j^{\text{med}}(Y) := \log_2 \left(\text{Median} \left\{ d_{j,k}^2, k = 1, \dots, N_j \right\} \right) - h_{N_j}, \quad (2.12)$$

where

$$h_{N_j} := \mathbb{E} \log_2 \left(\text{Median} \left\{ Z_k^2, k = 1, \dots, N_j \right\} \right) - \log_2 m(Z^2), \quad (2.13)$$

and where $Z, Z_k, k = 1, \dots, N_j$ are independent standard normal random variables. In (2.12) and (2.13), by “Median” we denote the sample median and by $m(\xi)$, we denote the median of the distribution of the random variable ξ . The term h_{N_j} serves as a first-order bias correction term and can be computed in practice by using Monte Carlo simulations. Observe that for large N_j , the term h_{N_j} is negligible. We will use the quantity

$$V_j := \text{Var} \left(\log_2 \text{Median} \left\{ Z_k^2, k = 1, \dots, N_j \right\} \right)$$

to estimate the variability of the statistic $S_j^{\text{med}}(Y)$. As with the term h_{N_j} , we compute V_j by using Monte Carlo methods. There are, at this point, no theoretical results about the properties of V_j and those of the median based spectrum S_j^{med} .

2.3 Applications: the curse of non-stationarity

The wavelet estimator often, in practice, may be the preferred estimator for the Hurst parameter. In particular, for non-contaminated data, it is essentially comparable to some of the best estimators such as the local Whittle estimator (Robinson (1995)). When the data are contaminated with smooth slowly varying trends, the wavelet estimator continues to work well, whereas the local Whittle estimator fails (see, for example, Stoev, Taqqu, Park and Marron (2004)).

On the other hand, as illustrated in Stoev, Taqqu, Park and Marron (2004), the wavelet estimator can be sometimes misleading. For example, high-frequency periodic deterministic components in the data affect the shape of the wavelet spectrum. Furthermore, non-stationarity effects such as abrupt shifts in the mean typically yield a *steep* wavelet spectrum and result in overestimating the Hurst parameter. These limitations suggest that, in practice:

(i) the wavelet estimator should not be used blindly, without a careful examination of the wavelet spectrum.

(ii) data should be examined for severe non-stationarity such as shifts in the mean; in particular when very long datasets are being analyzed, one should expect to encounter local non-stationarity and/or trends.

Consider, for example, Figure 1. It displays an Internet traffic trace (time series) collected at the University of North Carolina at Chapel Hill (UNC). The time series involves packet counts of

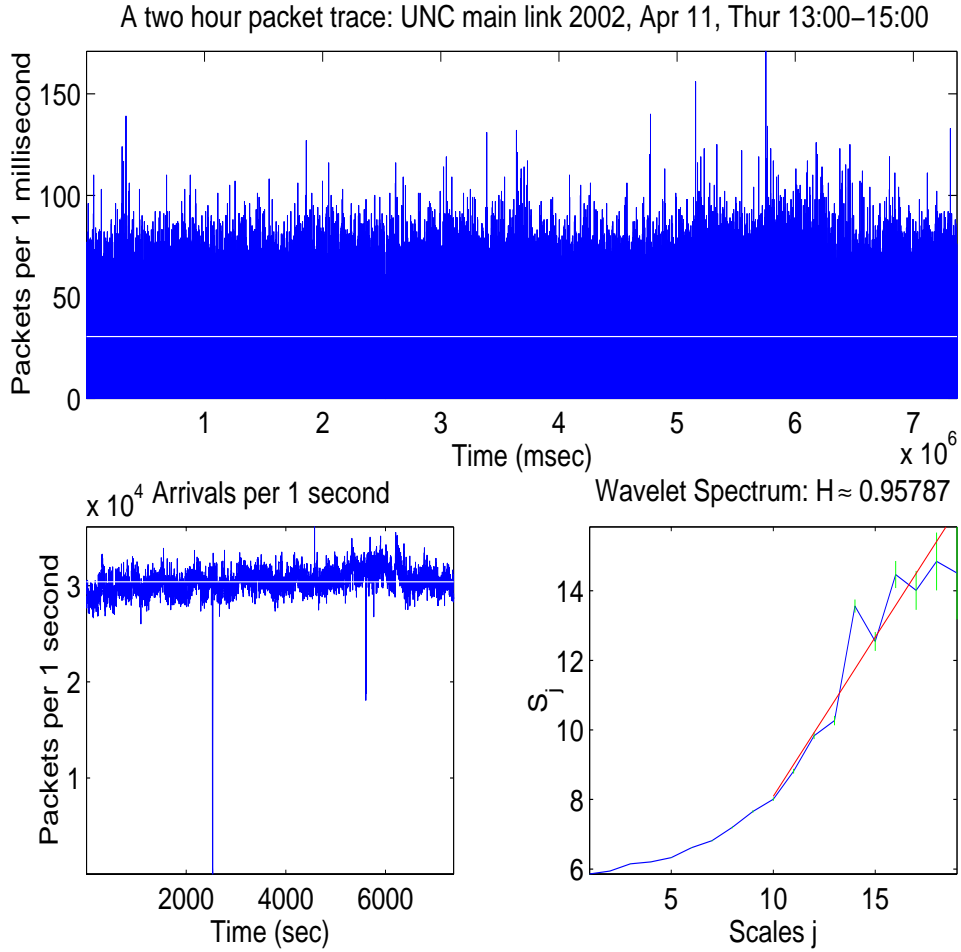


Figure 1: The top plot shows one two-hour trace of the number of packets arriving on a link per 1 millisecond time intervals. The data was collected at the UNC main link on Thursday, April 11, 2002 from about 13:00 to 15:00 o'clock. The mean is shown in white. The bottom-left plot shows the packet arrivals aggregated to 1 second time intervals. Observe the large drop in the traffic rate at time about 2500 sec. The wavelet spectrum is shown in the bottom-right plot. The Hurst parameter, estimated over the range of scales $10 \leq j \leq 20$, based on the slope of the fit line shown, is about $\hat{H} = \hat{H}_{[10,20]} \approx 0.958$. Note the large variability in the top portion of the spectrum, which suggests that a global Hurst exponent may not be meaningful.

the IP (Internet Protocol) packets in the Internet traffic coming into the UNC campus network. They were measured every 1 millisecond, for the duration of about two hours, from 13:00 to 15:00 o'clock on Thursday, April 11, 2002. The data has been processed from logs of IP packets by members of the DIstributed and Real-Time (DIRT) systems lab at the Department of Computer Science of UNC and is freely available from http://www-dirt.cs.unc.edu/unc02_ts/. The file name is `2002_Apr_11_Thu_1300.7260.sk1.1ms.B.P.ts.gz`.

The wavelet spectrum of the trace shown in this figure is consistent with long-range dependence (see, Hernández-Campos, Le, Marron, Park, Park, Pipiras, Smith, Smith, Trovero and Zhu (2004)). The variability in the spectrum at large scales, however, indicates inconsistency with the classical fractional Gaussian noise type of models. This casts doubt on a *global* analysis where one analyzes the long-range dependence in terms of a “global” Hurst parameter. In the following sections, we introduce practical tools for the estimation of the *local* Hurst parameter and we illustrate them by using the trace shown in Figure 1. These tools address the points (i) and (ii), above, by providing means for visualization of the local dependence structure in the data.

3 An animation tool to visualize the local dependence

We describe here an animation tool, that gives the opportunity to researchers to explore visually local dependence structures. The tool subdivides the data in windows and as the focus moves from one window to the next, one can clearly see the changes in the corresponding local wavelet spectrum. This tool, although simple, is quite informative and useful.

To describe the tool, consider a time series $Y(k)$, $k = 1, \dots, N$, choose an initial window size $w < N$ and divide the time series Y into $[N/w]$ non-overlapping time series Y_r , $r = 1, \dots, [N/w]$, where Y_r is the time series corresponding to the r th window. The first $[N/w] - 1$ windows are of size w and the last one is of size $N - w([N/w] - 1) \geq w$. Compute the wavelet spectrum of the time series Y within each window and obtain a matrix S of dimensions $(J \times [N/w])$, where $J = J(w) < \log_2(w)$ equals the number of available dyadic scales in each of the windows. As in (2.6), the (j, r) th element of the matrix S is defined as

$$S_j(r) := \log_2 \left(\frac{1}{N_j} \sum_{k=1}^{N_j} d_{j,k}^2(Y_r) \right) - g_{N_j}(j), \quad (3.1)$$

where $d_{j,k}(Y_r)$, $k = 1, \dots, N_j$ denote the wavelet coefficients of the time series Y_r corresponding to the r th window.

In Figure 2, we present one frame of the animation tool, which gives a preliminary view of the local dependence of the data. For each window $r = 1, \dots, [N/w]$, on the top-left plot we display the wavelet spectrum. The vertical segments indicate 95% confidence intervals for the statistics $S_j(r)$, $j = 1, \dots, J(w)$ of the current r th window, $r = 1, \dots, [N/w]$. On the bottom-left plot therein, we display a color (gray-scale) diagram of the entire matrix of local wavelet spectra and a vertical cursor focusing on the color-coded values of the column vector $(S_j(r))_{j=1}^J$ for the current frame.

The top part of this color diagram corresponds to the region of fine scales j of the wavelet spectra $S_j(r)$, $r = 1, \dots, [N/w]$ and the bottom part corresponds to the coarse scales j of the

spectra, respectively. A perfectly linear wavelet spectrum would correspond to evenly distributed colors in the columns of the color diagram.

The bottom-right plot of Figure 2 shows the local mean of the data, that is, the average of the time series Y per window. The vertical cursor indicates the position of the current frame. The top-right plot shows local wavelet estimates of the Hurst parameter, based on the wavelet spectrum within each window. We obtain these estimates by using a weighted linear regression over a set of scales j_1, \dots, j_2 , $j_1, j_2 \in \{1, \dots, J(w)\}$. The scales $j_1 < j_2$ are chosen by the user. As the frame moves, a vertical bar is presented indicating a 95% confidence interval for the estimated $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$.

This simple initial step of local analysis of the data can be very useful in many circumstances. It has the advantage of displaying in a succinct way the local dependence structure of the data as seen through the wavelet spectrum lens. One can clearly observe, and to an extent quantify, changes in the structure of the dependence. In fact, one can obtain local estimates of the Hurst parameter, which are often more meaningful and reliable than a single global one (compare Figures 1 and 2).

This view of the data, however, depends on two key choices:

- (i) the *size of the window* w and
- (ii) the *range of scales* $[j_1, j_2]$ used to estimate the local Hurst parameter $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$.

The next tool, described in the following Section 4, addresses these two issues. It also provides additional methods for visualization and statistical inference of the local wavelet spectrum.

4 A tool for the Local Analysis of Self-Similarity (LASS)

Here, we introduce a wavelet-based tool for the local analysis of the Hurst parameter in data. This tool was designed to facilitate the analysis of long datasets, which may exhibit non-stationarity in some regions. We illustrate the tool by using the Internet traffic trace displayed in Figure 1, and by using simulated fractional Gaussian noise.

4.1 Steps 1, 2 and 3: Visualizing the local self-similar scaling

When using the wavelet estimator of the Hurst exponent H of a LRD time series, it is crucial to choose sufficiently large starting dyadic scales j_1 . On large (coarse) scales, the long-range dependent time series become approximately self-similar with self-similarity parameter H . The wavelet spectrum at small or high-frequency scales, however, may not be linear or may not have the same slope, due to the specific short-term dependence structure of the time series. When estimating the Hurst parameter one chooses, in practice, the widest possible range of large scales, where the wavelet spectrum *is approximately linear*. Veitch, Taqqu and Abry (2003) proposed an automatic procedure for choosing the scales j_1 and j_2 . Here we present an alternative, graphical tool, to choose the range of scales where self-similar scaling is present. This visualization also provides the user with statistical confidence related to a choice of scales.

As in Section 3, above, we fix a window size w and compute the matrix S of the local wavelet spectra of the data (see (3.1)).

- *Step 1*

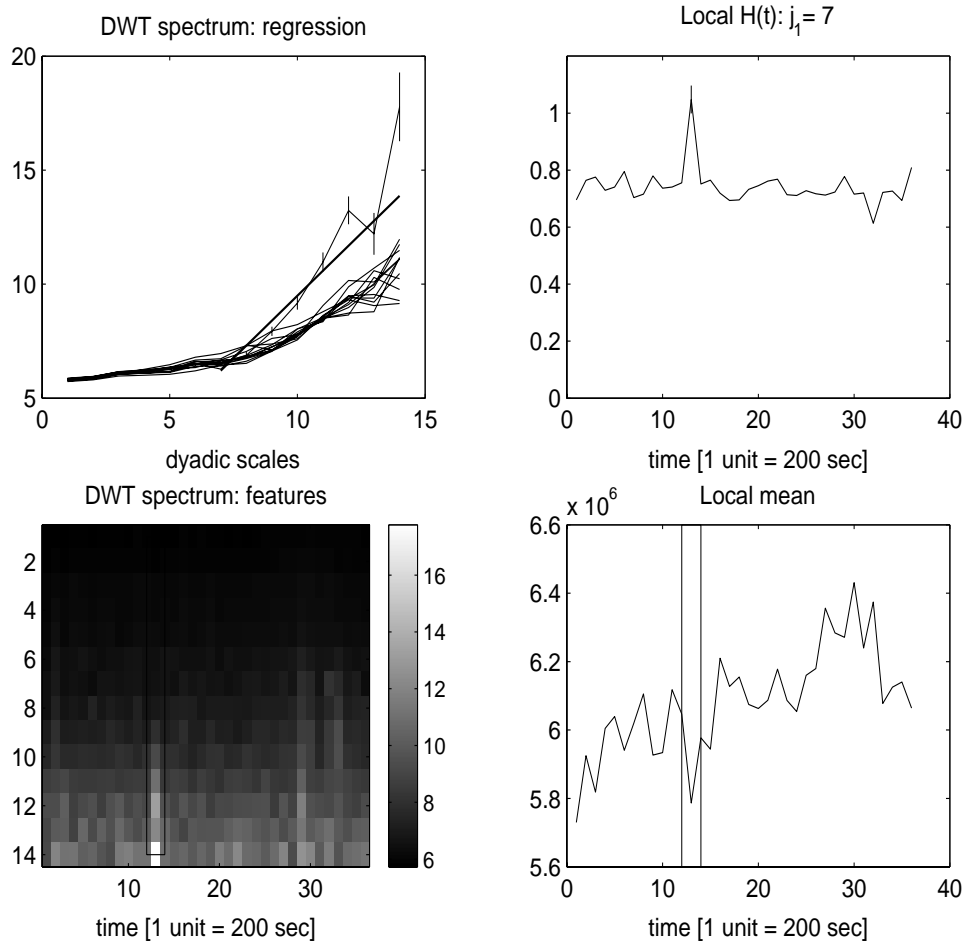


Figure 2: This figure represents one frame of an animation view of the local dependence structure of the Internet traffic trace shown in Figure 1 (number of packets per 1 ms). The top-left plot shows the wavelet spectrum for the current frame, the bottom-left plot shows color-coded values of all the wavelet spectra for the whole data set, the top-right plot shows the local estimates $\hat{H}(t)$ and the bottom-right plot presents the local mean of the data. The window used here is $w = 200\,000$ observations, which corresponds to 200 sec of traffic. This frame encompasses an extreme single drop in the data. The corresponding estimated value of the local Hurst exponent $\hat{H}(t)$ is about 1.12. Theoretically, the range of values for the Hurst exponent of a second-order stationary process is $(0, 1)$. The animation indicates that, with the exception of a few such fluctuations, the dependence structure of the presented trace is *locally* consistent with that of FGN on a wide range of coarse scales.

First, we set $j_1 = 1$ and $j_2 = J(w)$ and for each window $r = 1, \dots, [N/w]$, we estimate the local Hurst parameter $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ from all available wavelet scales (see the top-plot of Figure 3). We do so by using Relation (2.9) with weights $v_j \propto \sqrt{N_j}$, $j = 1, \dots, J-1$, where N_j is the number of wavelet coefficients used to obtain the statistic $S_j(r)$ (see (2.6)). While other choices of the weights v_j are possible, this choice reduces the sample variance of the estimators $\hat{H}(r)$, by putting more weight on statistics $S_j(r)$ with lower variance (for more details, see Abry and Veitch (1998)).

The bottom plot of Figure 3 displays statistically significant deviations from the estimated local Hurst parameters. Namely, the r th column in this plot corresponds to the vector of statistics $\hat{H}_{[j, j+1]}(r)$, $j = 1, \dots, j_2 - 1$, where, as in (2.8), $\hat{H}_{[j, j+1]}(r) = (S_{j+1}(Y_r) - S_j(Y_r))/2 + 1/2$. The j th cell in the k th column of the plot is colored blue (or black in gray-scale), when $\hat{H}_{[j, j+1]}(r)$ is *above* a 95% confidence interval about the estimate $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$; it is colored in red (or white) when $\hat{H}_{[j, j+1]}(r)$ is *below* this confidence interval; and it is colored in purple (or gray) when the estimate $\hat{H}_{[j, j+1]}(r)$ is *within* the confidence interval. The LASS tool offers two ways to compute the confidence interval:

(a) by assuming that the statistics $\hat{H}_{[j, j+1]}(r)$, $j = j_1, \dots, j_2 - 1$ are independent and Gaussian with variances $\text{Var}(S_{j+1}(r)) - \text{Var}(S_j(r))$.

(b) by using the sample covariance matrix of the vectors $(\hat{H}_{[j, j+1]}(r))_{j=j_1}^{j_2-1}$, $r = 1, \dots, [N/w]$.

The dependence of the statistics $\hat{H}_{[j, j+1]}(r)$, $j = 1, \dots, J(w) - 1$, is hard to evaluate, in particular when no model for the time series Y is imposed. In practice, however, as shown in Figure 4, the differences $\hat{H}_{[j, j+1]}(r)$, $j = 1, \dots, J(w) - 1$ are essentially *uncorrelated*. Their joint distribution can be also well-approximated by a multivariate normal distribution since the wavelet coefficients $\{d(j, k)\}_{j=1}^J$, $k \in \mathbb{Z}$ involved in the statistics $S_j(r)$ are weakly dependent (see, Bardet *et al.* (2000) and also Pipiras, Taquq and Abry (2001)). This justifies the assumptions in (a) above.

Although the time series can be non-stationary, we assume, for the purpose of computing the sample covariance matrix of \hat{H} that the estimates $\hat{H}_{[j, j+1]}(r)$ are stationary in r . This sample covariance matrix of the vectors $(\hat{H}_{[j, j+1]}(r))_{j=j_1}^{j_2-1}$, $r = 1, \dots, [N/w]$ can be a very good first approximation of the covariance structure of the local wavelet spectra. The choice in (b), which is data driven, is typically in close agreement with the choice in (a).

By taking into account the variability of the statistics $\hat{H}_{[j, j+1]}(r)$ on different scales j , the bottom plot of Figure 3 indicates whether the local wavelet spectra of the data scales linearly. It also shows statistically significant deviations from linear scaling. It can be used as a diagnostic tool for the estimation of the local Hurst parameter. When the plot displays many *red* (*white*) and *blue* (*black*) patches, the statistics $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ may not be reliable estimates of the local scaling exponents. One should perhaps change the choice of the scales j_1 and j_2 in order to improve the estimators $\hat{H}(r)$. A strategy for this is:

- *Step 2*

Based on the results of *Step 1* one may choose to focus on a different choice of scales j_1 and j_2 and to repeat the same analysis. In *Step 2* of the tool, by default, we set $j_2 = J(w)$ and let the user choose a new value of j_1 . In Figure 5 we show what happens when $j_1 = 7$ and $j_2 = J(w) = 15$, that is, by focusing on the largest scales of the wavelet spectra.

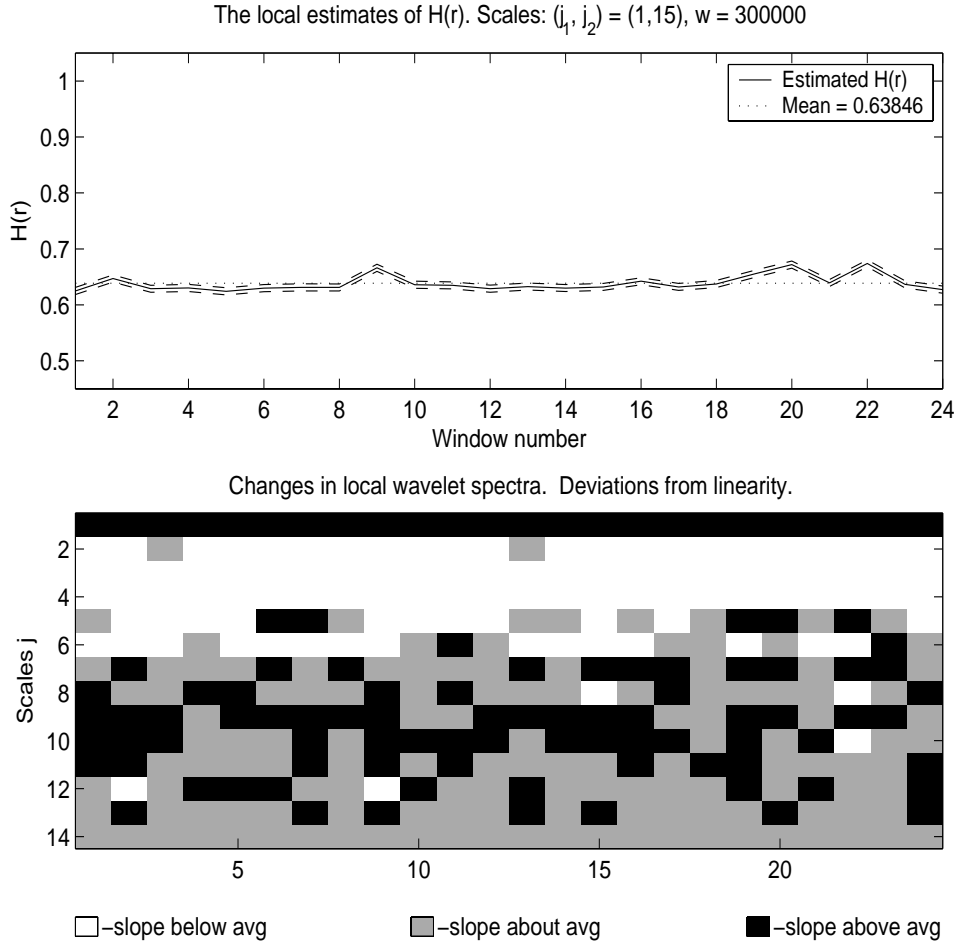


Figure 3: This figure displays the output of *Step 1* in the LASS tool applied to the two-hour Internet traffic data set shown in Figure 1 (number of packets per 1 millisecond). We used a *window size* $w = 300\,000$, corresponding to 5 minutes of traffic, Daubechies mother wavelet with 3 zero moments and *all available* scales $j \in [j_1, j_2] = [1, 15]$ to estimate the local Hurst parameters. The top plot displays the estimates $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ of the local Hurst parameter and a band of their 95% confidence intervals. The mean of the local Hurst estimates (in r) is indicated by the dotted line. The bottom plot visualizes (as explained in the text) the deviations from linearity in the local wavelet spectra. The presence of many *red* (*white*) and/or *blue* (*black*) patches indicates that the choice of scales j_1 and j_2 is not suitable and that these Hurst parameter estimates may not be meaningful (see Figure 5).

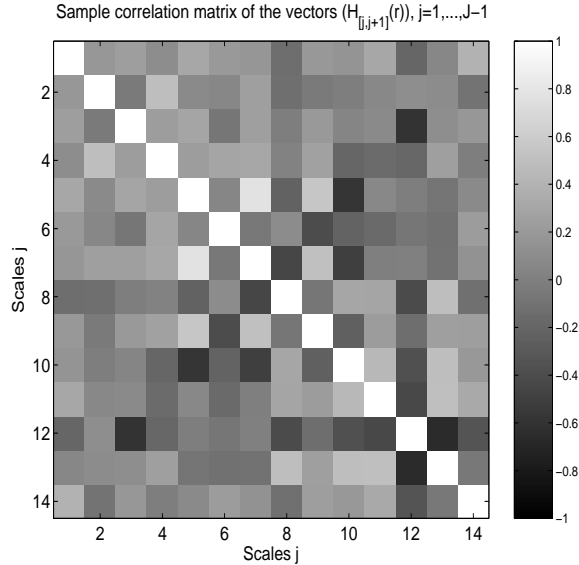


Figure 4: This figure displays a color-coded diagram of the sample correlation matrix of the statistics $\{\hat{H}_{[j,j+1]}(r)\}_{j=1}^{J-1}$ involved in Figure 3 above. Here $r = 1, \dots, [N/w]$ ($= 1, \dots, 24$), where the window size w equals 300 000. Observe that the statistics $\hat{H}_{[j,j+1]}(r)$ appear to be weakly correlated in j .

Observe that now fewer windows (columns of the bottom plot) show patches of *red* (*white*) and *blue* (*black*) as compared to the lower plot in Figure 3. The estimates $\hat{H}(r)$ of the local Hurst parameters have also changed. These estimates may be viewed as more reliable than the ones obtained in *Step 1*.

The bottom plot of Figure 5 indicates interesting non-stationary features in the local dependence structure of this data for window numbers $r = 9$ and 19 (corresponding to times t about 45 min and 100 min, respectively). These non-stationarities are also reflected in the top plot in Figure 5. (These features can also be seen in Figure 2.)

- *Step 3*

In this step, the tool presents a summary plot of the estimators of the local Hurst parameters in *Steps 1* and *2*. It also displays the local sample mean and local sample standard deviations of the time series, computed within each window frame. For brevity, we do not include the corresponding figure.

4.2 Step 4: The influence of the window size

In this step we explore the influence of the choice of the window size on the local estimates of the Hurst parameter. In the top-plot of Figure 6, we display a color diagram of local estimates $\hat{H}_{\text{interp}}(t)$ of the Hurst parameter, for several values of the window size w . The vertical axis corresponds to values of w . To be able to compare different window sizes, the horizontal axis of this plot is “time”. Since for different window sizes w , the estimate $\hat{H}(r) = \hat{H}_{[j_1,j_2]}(r)$ of *Steps*

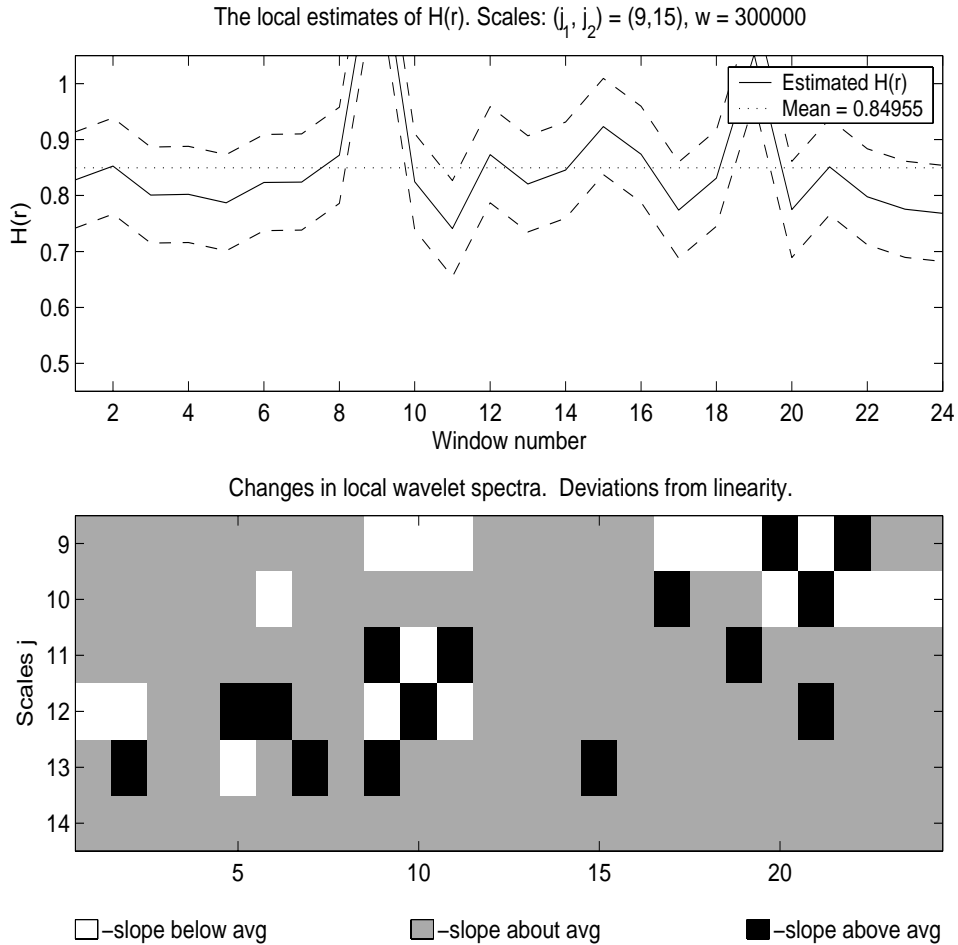


Figure 5: This figure contains the output of *Step 2* of the LASS tool, applied to the Internet traffic trace in Figure 1. As in Figure 3, the top plot displays estimates $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ of the local Hurst parameter, their point-wise 95% confidence intervals and their average over the window number parameter r . Now, however, to estimate the local Hurst parameters, we used the range of scales $[j_1, j_2] = [9, 15]$. Consequently, the bottom plot in Figure 5 is different from the bottom plot in Figure 3. There are relatively fewer *red* (*white*) and *blue* (*black*) patches in the bottom plot of Figure 5, which indicates that the local Hurst estimates displayed on the top plot therein are more *reliable* than those in Figure 3. Observe that the large fluctuations (at $r = 9$ and $r = 20$) in the estimates $\hat{H}(r)$ in Figure 5 correspond to significant deviations (*red* (*white*) and *blue* (*black*) patches) from linearity in the local wavelet spectra for the corresponding windows.

1 and 2 corresponds to different time locations $t \propto wr$, we set

$$\widehat{H}_{\text{interp}}(t) = \widehat{H}(r), \quad \text{for } C(r - 1/2)w < t \leq C(r + 1/2)w, \quad (4.1)$$

with $C = 1/w_0$, where w_0 is the minimum of the involved window sizes so that when $w = w_0$, $\Delta t = \Delta r = 1$. The correspondence between colors and values of $\widehat{H}_{\text{interp}}(t)$ is shown on the vertical color-bar on the top-plot of Figure 6.

The bottom plot of Figure 6 displays an overlay plot of the estimators $\widehat{H}_{\text{interp}}(t)$, one for each window size $w = 50\,000(50\,000)500\,000$. The average of these plots is displayed in bold. The estimates of $\widehat{H}_{\text{interp}}(t)$ were obtained by using the same starting scale $j_1 = 9$ and all available larger scales. This choice of j_1 corresponds to the one in *Step 2* (see Figure 5, above). Note that the estimates $\widehat{H}_{\text{interp}}(t)$ for small window sizes (e.g. $w = 50\,000$ and $100\,000$) have greater variability than those for larger window sizes (e.g. $w = 500\,000$). This is due to the fact that fewer wavelet coefficients are involved in the statistics $S_j(r)$ used to obtain $\widehat{H}_{\text{interp}}(t)$.

For the Internet data, displayed in Figure 1, note that in the top plot of Figure 6, looking along vertical lines, one color tends to predominate over many different window sizes. Thus, the resulting estimates of the local Hurst parameter are rather consistent, over all window sizes we used here. Note also the change in the pattern of the estimates around times $t = 50$ and 110 , providing another view of the non-stationarity observed in Figures 2, 3 and 5.

4.3 Additional features and options

In Sections 4.1–4.2, above, we described a tool for the local analysis of self-similarity. We did so by using only the basic default options. We will now list the tool’s other available options, which may provide additional insights into the local dependence structure of the data. Some of these options are illustrated in Section 5.

- *Visualization of the extreme fluctuations of $\widehat{H}(r)$*

The LASS tool involves an option which provides an additional view on the results of *Step 4*. It is based on the assumption that the estimates $\widehat{H}(r) = \widehat{H}_{[j_1, j_2]}(r)$, $r = 1, \dots, [N/w]$ are essentially uncorrelated and stationary. For each fixed window size w involved in *Step 4* (corresponding to a row on the top plot of Figure 6) using the sample $\widehat{H}(r)$, $r = 1, \dots, [N/w]$, it displays a 95% confidence interval for the mean of the $\widehat{H}(r)$ s based on a normal approximation. As in Figure 6, two plots are provided. The bottom plot displays the confidence intervals as a function of the window label and the top plot, which has the same axes as the top plot in Figure 6, displays color-coded extreme fluctuations of the local Hurst exponents. For a given window size w (indicated by its label) and location r (time), a cell is colored *blue* (*black*) if the local Hurst exponent estimate falls above its confidence interval, *red* (*white*) if it falls below its confidence interval and *purple* (*gray*) if it falls within the confidence interval. This color-coded view of the local Hurst estimates obtained on *Step 4* can be useful to localize extreme fluctuations in the local dependence structure of the data simultaneously for a set of window sizes.

- *Robust wavelet spectrum*

One can repeat the analysis on *Steps 1–4* by using the median-based robust wavelet spectrum defined in (2.12). In Section 5, below, we illustrate some advantages of this option.

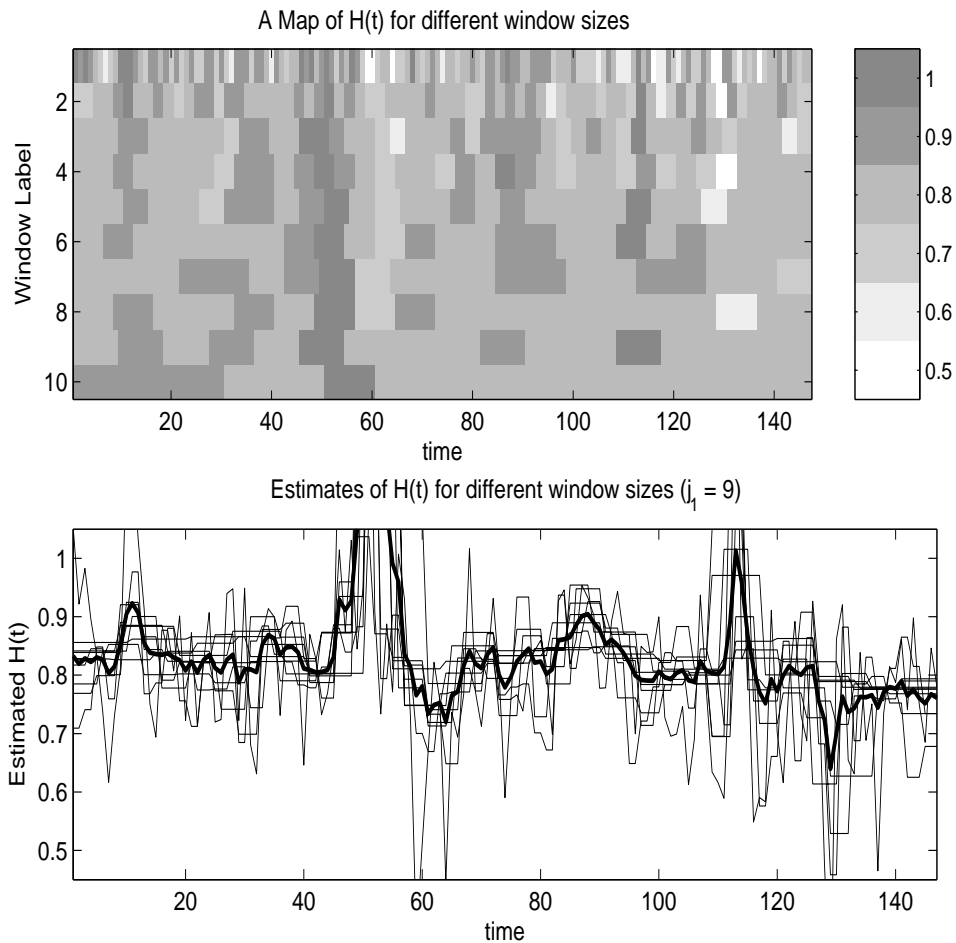


Figure 6: This figure displays the output of *Step 4* in the LASS tool, applied to the Internet data set in Figure 1. The top plot displays the color-coded values of the local Hurst estimates $\hat{H}_{\text{interp}}(t)$ in (4.1) for 10 window sizes: $w = 50\,000$ (50 000) 500 000 indicated by their window label 1 (1) 10. We obtained these estimates by using all available scales j greater than or equal to $j_1 = 9$ and used Daubechies wavelets with 3 zero moments. This plot visualizes the role of the window size in the estimation of the local Hurst exponents. Observe that, looking along vertical lines, one color tends to predominate over many different window sizes. The estimates $\hat{H}_{\text{interp}}(t)$ here are thus relatively robust with respect to the choice of the window size. In particular, the features observed at $r = 9$ and 19 in Figure 5 appear for all window sizes in the top plot of Figure 6 (note the vertical stripes at $t = 50$ and $t = 110$). The bottom plot displays an overlay plot of the estimates $\hat{H}_{\text{interp}}(t)$, for each window size $w = 50\,000$ (50 000) 500 000. Their average is displayed in bold.

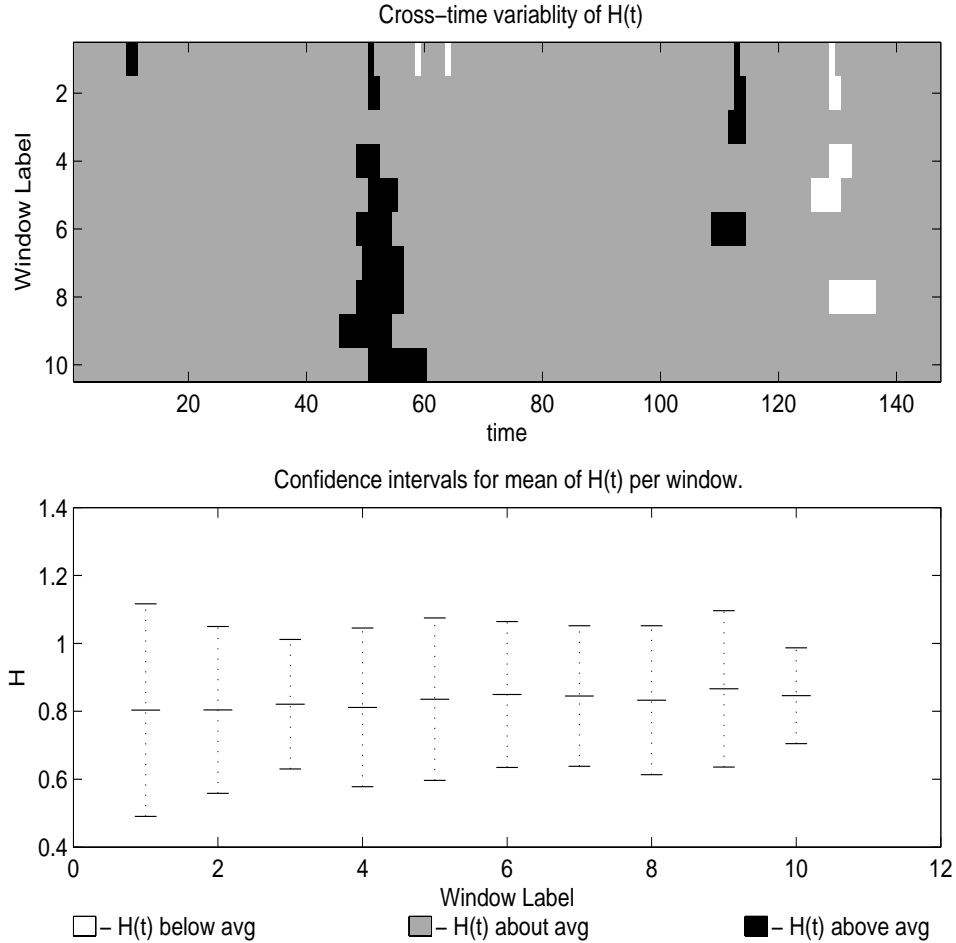


Figure 7: This figure visualizes the extreme fluctuations in the local Hurst estimates obtained on *Step 4* in the LASS tool (see Figure 6). The bottom plot displays 95% confidence intervals for the means of the estimates $\hat{H}_{\text{interp}}(t)$ displayed on the top plot of Figure 6. The horizontal axis corresponds to the window label (i.e. to the vertical axis of the top plot in Figure 6). The top plot visualizes the extreme outliers in the local estimates of the Hurst parameter which fall out of the confidence intervals in the bottom plot. The vertical axis of the top plot corresponds to the window label and the horizontal axis to time. For a given window size w (row) and a location t (time), a cell is colored in *blue/red* (or *black/white*, in gray-scale, respectively), if the estimate $\hat{H}_{\text{interp}}(t)$ falls *above/below* its corresponding confidence interval shown on the bottom plot. Observe that the fluctuations in $\hat{H}_{\text{interp}}(t)$ about times $t = 50$ and 110 , appear to be significant for a wide range of window sizes w .

- “log” type wavelet spectrum

One can also use the wavelet spectrum S_j^{\log} , defined in (2.11), in place of the usual spectrum S_j . This option is particularly useful when dealing with very heavy-tailed time series.

- The regression weights

The estimators $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ are obtained by using (2.9). One choice of weights is $v_j \propto \sqrt{N_j}$, where N_j denotes the number of available wavelet coefficients. Another option is to use the sample covariance matrix of the statistics $(\hat{H}_{[j, j+1]}(r))_{j=j_1}^{j_2-1}$ for $r = 1, \dots, [N/w]$. Namely, suppose that $\hat{H}_{[j, j+1]}(r)$, $j = j_1, \dots, j_2 - 1$ are jointly normal with mean H and covariance matrix equal to the sample covariance matrix. Under this assumption, one obtains weights v_j , which yield the best linear unbiased estimator for H .

In practice, when the number of available windows $[N/w]$ is sufficiently large (e.g. > 30) this choice of weights may lead to estimates $\hat{H}(r)$ with lower sample variance. However, when the time series appears to possess severe non-stationarity, one should use the first option instead.

- Confidence intervals in Steps 1 and 2

In order to compute the confidence intervals in *Steps 1* and *2*, one can use either method **(a)** or method **(b)** described earlier. In Section 4.1, we used method **(a)**.

- Analysis of a process with stationary increments

In Section 2.2, we used wavelets to analyze a *stationary*, long-range dependent time series. When the data are observations of a non-stationary stochastic process with stationary increments, one can develop a similar wavelet spectrum based methodology for the estimation of the (asymptotic) self-similarity parameter (see, for example, Pipliras, Taqqu and Abry (2001) and Stoev and Taqqu (2003)). All of the LASS tool options, above, are available for data that can be modeled by a process with stationary increments. In such a case, the resulting wavelet spectra and the estimates of the local (asymptotic) self-similarity parameters are very similar to the corresponding wavelet spectra and the estimates of the local Hurst parameters for the increments of the data.

5 Simulation performance of LASS

We first display and comment on the output of the LASS tool for the ideal benchmark situation when the data set is the fractional Gaussian noise. For brevity, we show only part of the results. Then, we illustrate the benefits of the *robust* wavelet spectrum option of the LASS tool for the Internet data set studied in the previous section.

- LASS for the fractional Gaussian noise

We simulated a fractional Gaussian noise (FGN) time series $Y(k)$, $k = 1, \dots, N$ of length $N = 1\,000\,000$ with Hurst parameter $H = 0.8$, using the fast Fourier transform. Figure 8 displays *Step 2* of the LASS tool, applied to this FGN time series.

The choice $j_1 = 3$ improved considerably the linear “fit” of the local wavelet spectra as compared to the choice $j_1 = 1$ on *Step 1* of the tool. (For brevity, we do not include the figure

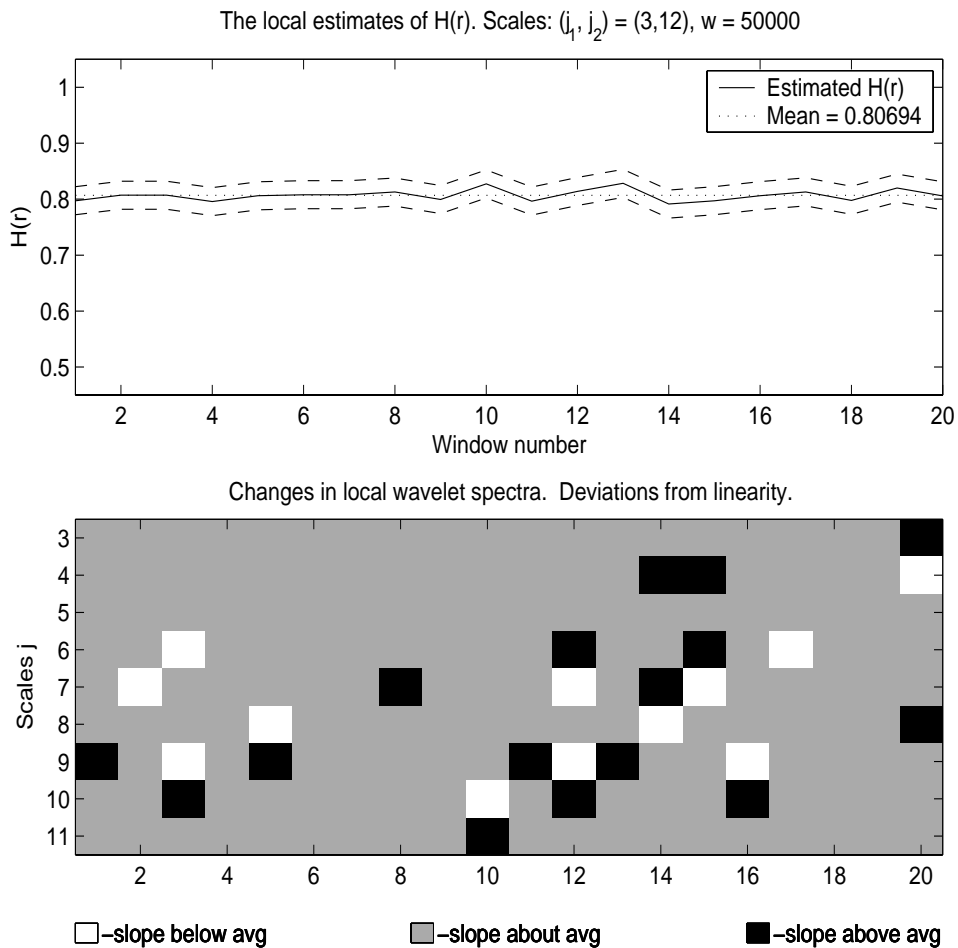


Figure 8: This figure displays *Step 2* in the LASS tool, applied to a simulated path of fractional Gaussian noise of length $N = 10^6$ and with Hurst parameter $H = 0.8$. We used Daubechies wavelets with 3 zero moments and weights $v_j \propto \sqrt{N_j}$ in (2.9) to estimate the local Hurst exponents, where N_j denotes the number of wavelet coefficients available on scale j . The bottom plot shows few *red* (*white*) and *blue* (*black*) patches, which suggests that the local estimates of the Hurst parameter (on the top plot) are reliable. (Contrast this plot with the one in Figure 3.) And indeed, all estimates $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$ are very close to the *true* value of $H = 0.8$.

from *Step 1.*) Observe that the estimate of the local Hurst parameter on the top plot of Figure 8 appears to be constant and very close to the true value of $H = 0.8$.

In contrast, for the Internet traffic trace, we had to choose a relatively large value $j_1 = 9$ in order to achieve a reasonably good linear fit in Figure 5. This is due to the intricate dependence structure of the Internet traffic on small time scales and the fact that fractional Gaussian noise can be used to model traffic only at sufficiently large time scales.

Figure 9 illustrates *Step 4* of the LASS tool for the FGN data. We used ten window sizes, as in Figure 6. Observe now that the top plot of Figure 9 is a lot more *homogeneous* than the one in Figure 6, which corresponds to the Internet traffic data. It clearly indicates no deviations from stationarity.

- *LASS with robust spectrum*

As illustrated above, the non-stationarity in Internet traffic clearly affects the classical wavelet estimator of the Hurst parameter (see also Stoev, Taqqu, Park and Marron (2004)). One may want to use, in practice, estimators of the *local* and *global* Hurst parameters, which are *robust* with respect to rare but significant non-stationarity artifacts. The *median-based* wavelet spectrum introduced in (2.12) can be used to obtain such robust estimators.

In Figure 10, we show *Step 2* of the LASS tool for the median-based spectrum in (2.12). Observe that, qualitatively, the bottom plots of Figures 5 and 10 are quite similar (in these two plots, we chose identical values for j_1 and for the window size). However, the estimates of the local Hurst parameter by using the robust spectrum, shown on the top plot of Figure 10, are now quite different from those on the top plot of Figure 5. Extremely large values of $\hat{H}(r) = \hat{H}_{[j_1, j_2]}(r)$, outside the range of $(0, 1)$, are no longer present.

6 Concluding remarks

Internet traffic traces possess long-range dependence on a wide range of time scales. Their dependence structure however is very intricate and can be obscured by non-stationarity effects. In practice, the classical fractional Brownian motion type models should be applied with care. In particular, a single parameter estimate of the Hurst long-range dependence exponent may not be meaningful or may not be sufficient to capture interesting network behavior.

On the other hand, it is important, in practice, to be able to understand the statistical structure of network traffic data sets, which can be very large. Here, we focused on estimating the Hurst parameter, locally in time, and on providing tools for visualization of the local dependence structure in large data sets. The local Hurst parameter estimates are more resilient to non-stationarity artifacts, provide a richer picture of the data and are often more meaningful than a single global one.

Acknowledgments

We would like to thank David Rolls, Arka Ghosh and Fred Godtliebsen for fruitful discussions. The Internet traffic data we use here have been processed from logs of IP packets by members of the DIstributed and Real-Time (DIRT) systems lab at UNC Chapel Hill

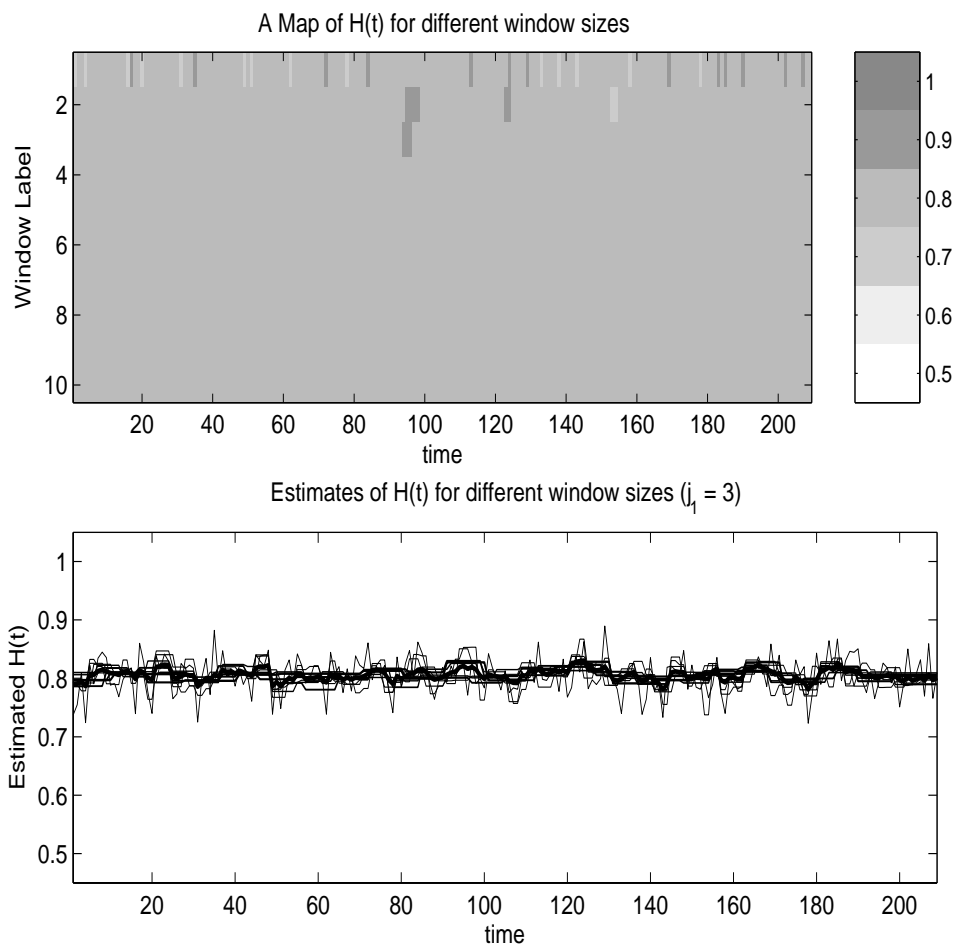


Figure 9: This figure displays *Step 4* in the LASS tool, applied to the simulated path of FGN, involved in Figure 8 with Hurst parameter $H = 0.8$. As in Figure 6, the top plot displays (color-coded) local Hurst estimates $\hat{H}_{\text{interp}}(t)$ for different values of the window size. Here, we used 10 window sizes $w = 5\,000$ (5000) 50 000, Daubechies wavelets with 3 zero moments and parameter $j_1 = 3$, for all values of w . That is, the local Hurst parameters were estimated by using all available scales j , greater than or equal to $j_1 = 3$. Observe that the top plot is a lot more homogeneous than the top plot in Figure 6, which corresponds to an Internet traffic trace. Consequently the overlay plot of the estimates $\hat{H}_{\text{interp}}(t)$, displayed on the bottom plot, shows no severe fluctuations and all estimates closely follow the theoretical value of $H = 0.8$.

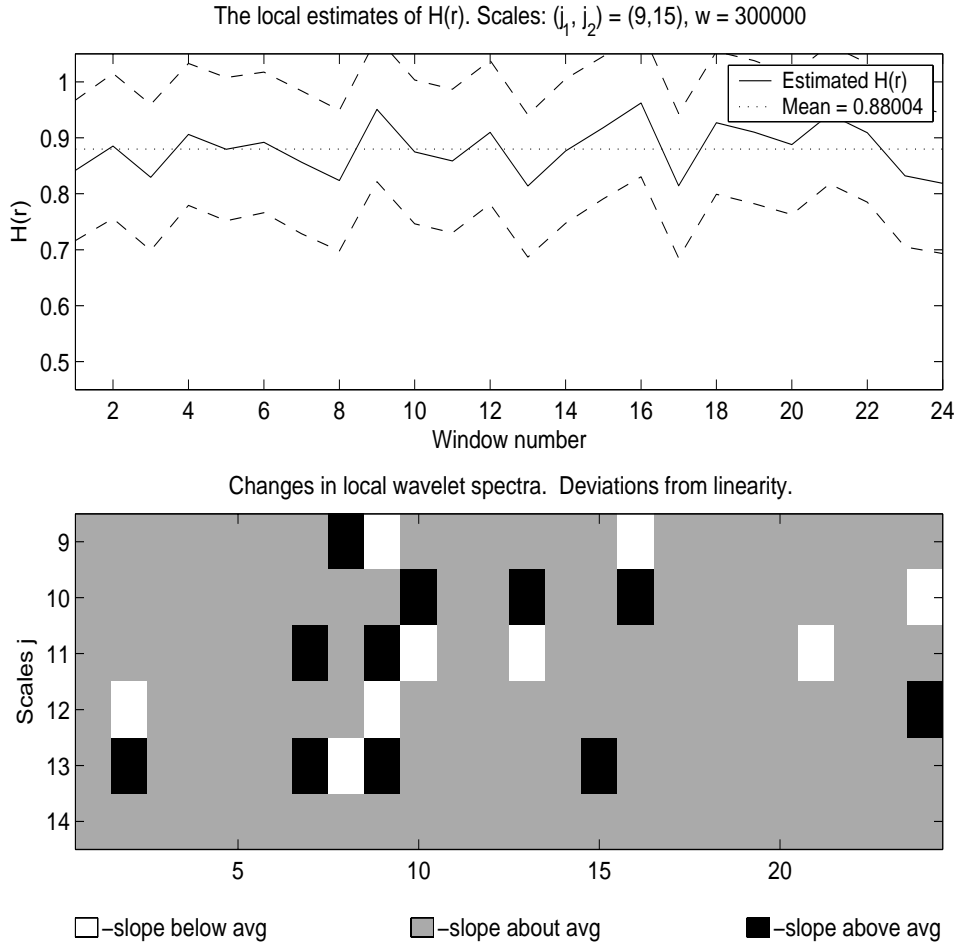


Figure 10: This figure displays *Step 2* in the LASS tool with the option of *median-based*, wavelet spectrum. We used the same Internet traffic data set involved in Figure 5 (shown on Figure 1). To be able to compare Figures 10 and 5, we used the same set of parameters: window size $w = 300\,000$, $j_1 = 9$, $j_2 = 15$ (maximum available wavelet scale) and Daubechies wavelets with 3 zero moments. Observe that now, the bottom plot of Figure 10 shows fewer *red* (*white*) and *blue* (*black*) patches than the bottom plot of Figure 5. The estimates of the local Hurst parameter in the top plot of Figure 10 have no extreme fluctuations above the theoretically admissible value of 1, in contrast with the local Hurst estimates in top plot of Figure 5. Note also that, on average, the local Hurst estimates based on the robust median spectrum are slightly greater than the local Hurst exponents in Figure 5.

(<http://www.cs.unc.edu/Research/dirt/>). Special thanks are due to Felix Hernández-Campos. This research was done while Murad S. Taqqu, George Michailidis and Stilian Stoev were visiting the University of North Carolina at Chapel Hill and the Statistical and Applied Mathematical Sciences Institute (SAMSI) at the Research Triangle Park. They would like to thank them for providing an excellent environment and inspiring working atmosphere. This research was also partially supported by the NSF Grant ANI-9805623 at Boston University and IIS-9988095 at the University of Michigan.

Appendix: Guide to the software tools

The software was written in MATLAB and may be obtained from the authors. We illustrate its use in this section. The examples, below, involve the exact set of parameters used to generate the figures in the paper.

- *Basic installation*

We suppose that all MATLAB script files of the tools are located in a directory `/home/me/lass/` (or a folder `C:\lass\` in MS Windows), for example. Under UNIX/Linux, one can use the command `unzip lass.zip`, where the file `lass.zip` is available on request from the authors.

After running MATLAB, make sure that the current directory is `/home/me/lass/` (or `C:\lass\`) or that this directory is in the *path* of MATLAB scripts. One can use the MATLAB command `addpath /home/me/lass/` to *add* the directory `/home/me/lass/` to the *path* of MATLAB scripts.

- *Loading a dataset*

Use the MATLAB command `load` to load a data set in MATLAB memory from a file. The file can be in either MATLAB binary format (i.e. MAT file) or in a suitable text format (i.e. space separated ASCII). For example, the command

```
>> data = load('2002_Apr_11_Thu_1300.7260.sk1.1ms.B.P.ts');
```

loads the Internet traffic data used in this paper, from the text file `2002_Apr_11_Thu_1300.7260.sk1.1ms.B.P.ts`, which is available from http://www-dirt.cs.unc.edu/unc02_ts/.

This data set (and hence the variable `data`) happens to have 2 columns. We will use the time series in the second column, namely `data(:,2)`.

- *Running the animation tool*

The command

```
>> animate3(data(:,2),struct('tau',0.001,'window',200000,'j1',7));
```

invokes the animation tool for the data in the column vector `data(:,2)` with a time unit of 0.001, a window size $w = 200\,000$ (corresponding to 200 sec of traffic) and parameter $j_1 = 7$.

- *Running the LASS tool*

The LASS tool is invoked by the command `lass` as follows:

```
>> lass(data(:,2),struct('window',300000,'windows',50000*[1:10]));
```


With this command, the window size w used in *Steps 1* and *2* is 300 000 and the window sizes used in *Step 4* are 50 000 to 500 000 with a step of 50 000. The output is Figure 3 using $j_1 = 1$ (*Step 1*) and then the user is prompted to choose another value of j_1 . The outputs are Figure 5 (*Step 2* using $j_1 = 9$), a figure for *Step 3* and Figure 6 (*Step 4*) as well as Figure 7.

The following defaults are used in both tools: Daubechies wavelets with 3 zero moments, initialize the Mallat's algorithm as proposed in Veitch, Taqqu and Abry (2000) and compute the Hurst parameters by using weighted regression where $v_j \propto \sqrt{N_j}$ (see (2.9)).

Typing `>> help animate3` or `>> help lass` yields detailed information about all the options.

References

- Abry, P. & Veitch, D. (1998), 'Wavelet analysis of long range dependent traffic', *IEEE Transactions on Information Theory* **44**(1), 2–15.
- Abry, P., Flandrin, P., Taqqu, M. S. & Veitch, D. (2000), Wavelets for the analysis, estimation and synthesis of scaling data, in K. Park & W. Willinger, eds, 'Self-Similar Network Traffic and Performance Evaluation', Wiley (Interscience Division), New York, pp. 39–88.
- Bardet, J.-M., Lang, G., Moulines, E. & Soulier, P. (2000), 'Wavelet estimator of long-range dependent processes', *Statistical Inference for Stochastic Processes* **3**, 85–99.
- Beran, J. (1994), *Statistics for Long-Memory Processes*, Chapman & Hall, New York.
- Daubechies, I. (1992), *Ten Lectures on Wavelets*, SIAM Philadelphia. CBMS-NSF series, Volume 61.
- Hernandez-Campos, F., Le, L., Marron, J.-S., Park, C., Park, J., Pipiras, V., Smith, F. D., Smith, R. L., Trovero, M. & Zhu, Z. (2004), Long range dependence analysis of Internet traffic, In preparation.
- Kim, M. & Tewfik, A. H. (1992), 'Correlation structure of the discrete wavelet coefficients of fractional Brownian motion', *IEEE Transactions on Information Theory* **38**(2), 904–909.
- Leland, W. E., Taqqu, M. S., Willinger, W. & Wilson, D. V. (1993), 'On the self-similar nature of Ethernet traffic', *Computer Communications Review* **23**, 183–193. Proceedings of the ACM/SIGCOMM'93, San Francisco, September 1993. Reprinted in *Trends in Networking – Internet*, the conference book of the Spring 1995 Conference of the National Unix User Group of the Netherlands (NLUUG). Also reprinted *Computer Communication Review*, **25**, Nb. 1 (1995), 202-212, a special anniversary issue devoted to "Highlights from 25 years of the Computer Communications Review".
- Mandelbrot, B. B. & Van Ness, J. W. (1968), 'Fractional Brownian motions, fractional noises and applications', *SIAM Review* **10**, 422–437.
- Mikosch, T., Resnick, S., Rootzén, H. & Stegeman, A. (2002), 'Is network traffic approximated by stable Lévy motion or fractional Brownian motion?', *The Annals of Applied Probability* **12**(1), 23–68.
- Park, K. & Willinger, W., eds (2000), *Self-Similar Network Traffic and Performance Evaluation*, J. Wiley & Sons, Inc., New York.
- Paxson, V. & Floyd, S. (1995), 'Wide area traffic: The failure of Poisson modeling', *IEEE/ACM Transactions on Networking* **3**, 226–244.
- Pipiras, V., Taqqu, M. S. & Abry, P. (2001), Asymptotic normality for wavelet-based estimators of fractional stable motion, Preprint.
- Robinson, P. M. (1995), 'Gaussian semiparametric estimation of long range dependence', *The Annals of Statistics* **23**, 1630–1661.

- Samorodnitsky, G. & Taqqu, M. S. (1994), *Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, London.
- Stoev, S. & Taqqu, M. S. (2003), ‘Asymptotic self-similarity and wavelet estimation for long-range dependent FARIMA time series with stable innovations’, *Journal of Time Series Analysis*. To appear (Preprint 2002).
- Stoev, S., Pipiras, V. & Taqqu, M. S. (2002), ‘Estimation of the self-similarity parameter in linear fractional stable motion’, *Signal Processing* **82**, 1873–1901.
- Stoev, S., Taqqu, M. S., Park, C. & Marron, J. S. (2004), Strengths and limitations of the wavelet spectrum method in the analysis of Internet traffic, Preprint.
- Taqqu, M. S. (1975), ‘Weak convergence to fractional Brownian motion and to the Rosenblatt process’, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **31**, 287–302.
- Taqqu, M. S. (2002), ‘The modeling of Ethernet data and of signals that are heavy-tailed with infinite variance’, *Scandinavian Journal of Statistics* **29**(2), 273–295.
- Taqqu, M. S. (2003), Fractional Brownian motion and long-range dependence, in P. Doukhan, G. Oppenheim & M. S. Taqqu, eds, ‘*Theory and Applications of Long-range Dependence*’, Birkhäuser, pp. 5–38.
- Taqqu, M. S., Willinger, W. & Sherman, R. (1997), ‘Proof of a fundamental result in self-similar traffic modeling’, *Computer Communications Review* **27**(2), 5–23.
- Veitch, D. & Abry, P. (1999a), A statistical test for the time constancy of scaling exponents, Preprint.
- Veitch, D. & Abry, P. (1999b), ‘A wavelet-based joint estimator of the parameters of long-range dependence’, *IEEE Transactions on Information Theory* **45**(3), 878–897.
- Veitch, D., Taqqu, M. S. & Abry, P. (2000), ‘Meaningful MRA initialization for discrete time series’, *Signal Processing* **80**, 1971–1983.
- Veitch, D., Taqqu, M. S. & Abry, P. (2003), ‘On the automatic selection of the onset of scaling’, *Fractals* **11**(4), 377–390.

Stilian Stoev
 Department of Mathematics and Statistics,
 Boston University,
 Boston, MA 02215
sstoev@bu.edu

Murad Taqqu
 Department of Mathematics and Statistics,
 Boston University,
 Boston, MA 02215
murad@bu.edu

Cheolwoo Park
 Statistical and Applied Mathematical Sciences Institute,
 19 T. W. Alexander Drive, P.O. Box 14006,
 Research Triangle Park, NC 27709-4006
cwpark@samsi.info

George Michailidis
 Department of Statistics,
 University of Michigan,
 459 West Hall
 Ann Arbor, MI 48109-1092
gmichail@umich.edu

J. S. Marron
 Department of Statistics,
 University of North Carolina,
 Chapel Hill, NC 27599-3260
marron@email.unc.edu