Long Range Dependence Analysis of Internet Traffic

Cheolwoo Park
Statistical and Applied Mathematical Sciences Institute
P.O. Box 14006
Research Triangle Park, NC 27709-4006
and Department of Statistics
University of Florida
103 Griffin/Floyd Hall - P.O. Box 118545
Gainesville, FL 32611-8545

Félix Hernández-Campos
Department of Computer Science
University of North Carolina
Chapel Hill, NC 25799-3175

Long Le
Department of Computer Science
University of North Carolina
Chapel Hill, NC 25799-3175

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

Juhyun Park
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

Vladas Pipiras
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

F. D. Smith
Department of Computer Science
University of North Carolina
Chapel Hill, NC 25799-3175

Richard L. Smith
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

Michele Trovero
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

Zhengyuan Zhu
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

September 18, 2004

Abstract

Long Range Dependent time series are endemic in the statistical analysis of Internet traffic. The Hurst Parameter provides good summary of important self-similar scaling properties. Here a number of different Hurst parameter estimation methods, and some important variations are compared. This is done in the context of a wide range of simulated, laboratory generated and real data sets. Important differences between the methods are highlighted. Deep insights are revealed on how well the laboratory data mimic the real data. Non-stationarities, that are local in time, are seen to be central issues, and lead to both conceptual and practical recommendations.

1 Introduction

The Internet has brought major changes to the work places, and even the lifestyles, of many people. It also provides a rich source for research problems at several levels of interest to statisticians. Two of these levels are web page content, and the mechanics of the traffic flows on network links. In this paper we study the latter. Background and motivation for the work done in this paper is given in Section 1.1.

A major goal of this paper is a careful comparison of Internet traffic models with simulated data, with laboratory generated data, and with real data. Because of the widely accepted Long Range Dependent (LRD) self-similar properties of network traffic (the history and reasons behind this are explained in Section 1.1), Hurst parameter estimation provides a natural approach to studying such models. In Section 2, a variety of Hurst parameter estimators are considered, including the Aggregated Variance method in Section 2.1.1, the Local Whittle approach in Section 2.1.2, and the Wavelet method in Section 2.1.3.

Detailed comparison of these methods is done in Section 3. Our comparisons are done in the context of an unusually large test bed of examples, described in Section 1.2. The number is such that even organizing the results is not a simple task. We addressed this issue through the construction of a summary web page, Le, Hernández-Campos and Park (2004). This has the added advantage of allowing interested readers to view the full results. In addition to all analyses appearing there, there are also links to all of the raw data sets. Here we report the most interesting lessons learned from these analyses.

The main lessons are that there are important differences between the various Hurst parameter estimators, and between the different types of data considered. While some often proposed models, such as Fractional Gaussian Noise (FGN), can fit the data well, quite troublesome non-stationarities are a very serious issue, which is studied in detail in Section 4. These lead to some recommendations for improved simulation and laboratory experiments, given in Section 4.4.

1.1 Background and Motivation

In the area of controlling Internet traffic flows, there is a wide range of open research problems for engineers, computer scientists, statisticians and probabilists. There are two fundamental issues that must be addressed; flow control, and congestion control. Flow control is necessary to ensure that the sending application does not send data at a rate exceeding the rate at which the receiving application can process it. Congestion control is necessary to ensure that the sending application does not send data at a rate exceeding the currently available transmission capacity at links along the network path from the sender to the receiver. If link capacity is exceeded, data are lost and must be retransmitted. The Transmission Control Protocol (TCP) is the most commonly used set of rules and algorithms used to transfer data on the Internet. It provides mechanisms for both flow control and congestion control. For a good introduction to the Internet, its protocols, and the issues in managing traffic flows, see the book by Kurose and Ross (2004). Unfortunately, engineering experience has shown that there are still a number of directions in which there is room for improvement in TCP. There is active research underway in this direction (for pointers to most of this research, see Floyd (2004)).

An important hurdle to research in congestion control is: which of the many new designs being developed should be used? Simulation is a natural tool for comparing designs, but developing simulators that effectively mimic Internet traffic also turns out to be a challenging problem. In particular, there are several generations of models that could be used. This motivates the statistical problem of goodness of fit of these various models to real Internet traffic, which is the main focus of this paper.

There are interesting parallels between the telephone network and the Internet. Both are global networks transporting large amounts of information between very diverse locations. Both are a concatenation of many pieces of equipment. But there are some important differences, which have a major impact on traffic modeling.

One difference is the handling of congestion. For the telephone network, a phone call essentially entails the sole use of a pair of wires from one party to the other. This approach is called “circuit switching” and requires a dedicated circuit path for each connected call even if neither party is actually using (talking on) the connection. Congestion in the network means there are no more circuits available for new calls which results in the familiar busy signal, and in the connection not being made. On the Internet, resources are much more effectively shared by splitting data transmissions into "packets" (small segments of the data being transferred). This approach is called "packet switching" and does not require network capacity to be dedicated to any one flow of packets. Instead, the capacity is shared on-demand as new packets arrive. Depending on the statistics of packet arrivals, several flows can concurrently share the same equipment because their packets' transmissions are interleaved with each other. There is no notion of busy signal, but the volume of packet arrivals in some interval of time can overwhelm the network's capability. When this happens, packets are lost, and need to be retransmitted. The value of TCP is in recognizing when such packet loss occurs, and in providing a mechanism for orderly retransmission of packets as well as adjusting the rate at which applications can send packets into a congested network.

An important statistical difference between the telephone network and the Internet comes in the distribution of the length of connections. The exponential distribution has provided a useful workhorse model for the telephone network, perhaps because human choice determines the duration of a phone call. But it has been shown in a number of places, see e.g. Paxson (1994), Garrett and Willinger (1994), Paxson and Floyd (1995), Crovella and Bestavros (1996) and Hernández-Campos, Marron, Samorodnitsky and Smith (2004), that the exponential distribution is very inappropriate for durations of Internet connections. For example there are very many connections which are much shorter than any phone call, and there are a few that are much longer than most people are interested in holding the phone to their ears.

Duration distributions show a simple way in which modelling for the two networks should be different. Another important difference comes from study of the traffic flow. The field of queueing theory has drawn much motivation from modeling traffic in the telephone network, for nearly a century. The standard assumptions are that phone calls are initiated according to a Poisson process, and the duration follows an exponential distribution. But both of these assumptions are unacceptably crude approximations for Internet traffic, as shown by Marron, Hernández-Campos and Smith (2004).

A major move away from standard queueing models was taken by a series of papers, including Paxson and Floyd (1995), Feldmann, Gilbert and Willinger (1998) and Riedi and Willinger (1999), that was based on the theory originally developed in other contexts by Mandelbrot (1969) and Taqqu and Levy (1986). Simple visual insight into such models comes from Figure 1. Flows of data over the Internet (more precisely a single HTTP, i.e. web browsing, file transfer), over a 225 second time interval are represented as horizontal lines, whose left endpoint indicates the start time, and whose right endpoint indicates the ending time. To avoid boundary effects, this time interval is chosen as the center of a four hour time window, on a Sunday morning in April 2001. The vertical coordinate of line segments is random, to allow good visual separation (essentially the jitter plot idea of Tukey and Tukey (1990)). Note that there are very many short connections, sometimes called "mice", and very few long connections, sometimes called "elephants". A simple visual way of seeing that this distribution is very far from exponential is to repeat the plot using segment lengths from the fit (using the sample mean) exponential distributions. This is not shown here to save space, but see Figure 5 of Marron, Hernández-Campos and Smith (2002), where the distribution has neither elephants nor mice, but many intermediate length line segments.

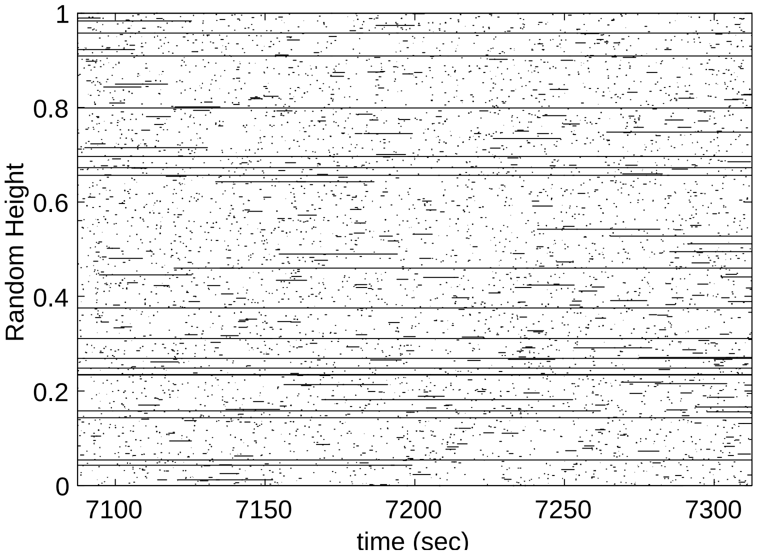

FIGURE 1: *Mice and Elephants plot, showing HTTP file transfers, with time shown as line segments. Random vertical height for visual separation.*

Many more such plots, together with some interesting and perhaps surprising sampling issues that are driven by the special scaling properties of these data may be found in Marron, Hernández-Campos and Smith (2002), and in even more detail at the corresponding web site.

Models for aggregated traffic, i.e. the result of summing the flows shown in Figure 1, are quite different from standard queueing theory when the distribution of the lengths are heavy tailed. Appropriate levels of heavy tails can induce long range dependence, with interesting scaling properties, as shown by the above authors. See also Resnick and Samorodnitsky (1999) for a clear exposition.

In this paper, aggregated traffic is studied via time series of packet counts over bin grids. Binned time series come from taking the timestamps of individual packets (the "atoms" that make up e.g. all of the line segments in Figure 1), and reporting the counts of the number of packets with timestamp values that fall between an equally spaced grid of bin boundaries in time (typically 1 millisecond bins).

1.2 Empirical Traffic Data Sets

There were two types of data sets derived from empirical data used for our analysis; traces of real Internet traffic and traces from laboratory experiments using synthetic traffic generators. Traces of real Internet traffic were obtained by placing a network monitor on the high-speed link connecting the University of North Carolina at Chapel Hill (UNC) campus network to the Internet via its Internet service provider (ISP). All units of the university including administration, academic departments, research institutions, and a medical complex (with a teaching hospital that is the center of a regional health-care network) used a single high-speed link for Internet connectivity. The user population is large (over 35,000) and diverse in their interests and how they use the Internet - including, for example, email, instant messaging, student "surfing" (and music downloading), access to research publications and data, business-to-consumer shopping, and business-to-business services (e.g., purchasing) used by the university administration. There are over 50,000 on-campus computers (the vast majority of which are Intel architecture PCs running some variant of Microsoft Windows). The core campus network that links buildings and also provides access to the Internet is a switched 1 Giga (billion) bits per second (Gbps) Ethernet.

We used a network monitor to capture traces of the TCP/IP headers on all packets entering the campus network from the Internet. All the traffic entering the campus from the Internet traversed a single full-duplex 1 Gbps Ethernet link from the ISP edge router to the campus aggregation switch. In this configuration both "public" Internet and Abilene (a backbone network operated by the Internet 2 Consortium to provide networking services to universities and to foster networking research) traffic were co-mingled on the one Ethernet link and the only traffic on this link was traffic to or from the Internet. The trace entry for each packet included an arrival timestamp (with 1 microsecond resolution and approximately 1 millisecond accuracy), the length of the Ethernet frame, and the complete TCP/IP header (which includes the IP length field). Each trace was processed offline to create a time series of the counts of packets arriving on the link in successive one millisecond intervals.

The traces were collected during selected two-hour tracing intervals over a one week period during the second week in April 2002, a particularly high traffic period on our campus coming just a few weeks before the semester ends. The two-hour intervals were chosen somewhat arbitrarily to have traces during the highest loads of the normal business day (e.g., 8:00 AM, 10:00 AM, 1:00 PM and 3:00 PM), during non-business hours when traffic volumes were the lowest (3:00 AM and 5:00 AM) or when "recreational" usage was likely to be high (7:30 PM and 9:30 PM). Both weekday and weekend times were included. In aggregate, these traces represent 40 hours of network activity and include information about 3.55 billion packets that carried 1.17 Tera (10^{12}) bytes of data. The average load on the 1 Gbps link during these two-hour tracing intervals ranged from a low of 2.7% utilization (Tuesday at 5:00 AM) to 9.1% (Thursday at 3:00 PM).

The synthetic traffic traces were obtained in a laboratory testbed network designed to emulate a campus network having a single link to an upstream Internet service provider (ISP). The intent, obviously, is to emulate an environment similar to the UNC network link. All traffic using the link is modeled as Web traffic where the Web clients are all located on the campus network and all the Web servers are located somewhere on the Internet beyond the upstream link. At one end of the laboratory link is a network of 18 machines that run instances of a synthetic Web-request generator each of which emulates the Web browsing behavior of hundreds of human users. At the other end of the link is another network of 10 machines that run instances of a synthetic Web-response generator that creates traffic to model the outputs from real web servers. In this paper and the web page, we refer to the Web request generators simply as "clients" and the Web response generators as "servers".

The link between clients and servers in the laboratory is a full-duplex 1 Gbps Ethernet link just like the real UNC link. So we can emulate packet flows that traverse a longer network path than the one in our laboratory, we use a software tool to impose extra delays on all packets traversing the network. These delays emulate a different randomly selected round-trip time for each client-to-server flow. The distribution of delays was chosen to approximate a typical range of Internet round-trip times within the continental U.S. The traffic load on the link is controlled by a parameter to the request-generating programs that specifies a fixed-size population of users browsing the web. The loads used in the experiments for the data presented here represent 2% , 5%, 8%, 11% and 14% of the link's capacity (roughly the same as the loads on the real UNC link). For example, a load of 11% is achieved by emulating over 20,000 active web users. The number of emulated users is held constant throughout the execution of each experiment.

The synthetic traffic is generated using a model derived from a large-scale analysis of web traffic by Smith, et al (2001) that describes how the HTTP 1.0 and 1.1 protocols are used by web browsers and servers. The model is quite detailed as it, for example, includes the use of persistent HTTP connections and distinguishes between web objects that are "top-level" (e.g., HTML files) and objects that are embedded (e.g., image files). The model is expressed as empirical distributions describing the variables necessary to generate synthetic Web workloads. The client and server programs create traffic by randomly sampling from these empirical distributions. The specific traffic generating programs are described in more detail in Le, et al (2003). The empirical distributions that have the most pronounced effects on generated traffic are the distribution of server response (file) sizes and user "think" times between requests. Each experiment was run for 120 minutes to ensure very large samples (over 10,000,000 request/response exchanges) but data were collected only during the middle 80 minutes (approximately) to eliminate startup effects at the beginning and termination synchronization anomalies at the end.

The motivation for using synthetic Web traffic in our experiments was the assumption that properly generated traffic would exhibit properties in the laboratory network consistent with those found in empirical studies of real networks, specifically, a long range dependent (LRD) packet arrival process. The empirical data used to generate our web traffic has heavy-tailed distributions for both user "think" times and response (file) sizes (see Smith, et al 2001). That our traffic generators used heavy-tailed distributions for both think times (OFF times) and response sizes (ON times), implies that the aggregate traffic generated by our large collection of sources (emulated users) should be LRD according to the theory developed by Willinger, Taqqu, Sherman and Wilson (1997).

2 Long Range Dependence and Hurst Parameter Approaches

This section contains an overview of the Hurst parameter, followed by a summary of some important methods of estimation, in Section 2.1.

Long range dependence is a property of time series that exhibit strong dependence over large temporal lags. For example, as mentioned above, we expect the data traffic analyzed in this work to be long range dependent because the "elephant" connections in Figure 1 affect traffic patterns over large time scales. Long range dependence can be formally defined in a number of essentially equivalent ways. Let $X_t, t \in \mathbb{Z}$, be a discrete time, second order stationary series (in our study, X_t is the time series of packet bin counts). The series $X_t, t \in \mathbb{Z}$, is called *long range dependent* (LRD, in short) if its covariance function

$$\gamma(t) = E\left(X_0 - EX_0\right)\left(X_t - EX_t\right) \sim c_\gamma |t|^{2-2H}, \quad \text{as } t \to \infty, \tag{1}$$

with

$$H \in (1/2, 1), \tag{2}$$

where $c_\gamma > 0$ is a constant. Condition (1) can be equivalently restated in the Fourier frequency domain as

$$f(\lambda) \sim c_f |\lambda|^{2H-1}, \quad \text{as } \lambda \to 0, \tag{3}$$

where $f(\lambda) = (2\pi)^{-1} \sum_k e^{-i\lambda k}\gamma(k), \lambda \in [-\pi, \pi]$, is the spectral density function of X_t and $c_f > 0$ is a constant. The parameter H in (2) is called the *Hurst parameter*. When H is larger, the temporal dependence is stronger because the covariance function $\gamma(t)$ decays more slowly at infinity. In all of these LRD contexts, the decay of $\gamma(t)$ is much slower than the exponential convergence typical of (short range) classical models, such as Auto-Regressive Moving Averages, see for example Brockwell and Davis (1996). For more information on LRD time series, see for example Beran (1994) and Doukhan, Oppenheim and Taqqu (2003).

For some purposes, other parametrizations can be more convenient. For example, motivated by the form of the exponent in (3), the related parameters $d = H - 1/2$ and $\alpha = 2d = 2H - 1$, are useful. Sometimes it is also useful to work with time series in continuous time, but in this paper we will only consider discrete time. Note also that LRD is often defined, more generally, by replacing the constants by slowly varying functions in (1) and (3). We use the constants here for simplicity, and also because commonly used methods to estimate H assume either (1) or (3).

On occasion, some estimates of H fall outside of the range $(1/2, 1)$. While Gaussian stochastic processes, with $H \geq 1$, exist, they are non-stationary, and the conventional mathematics of statistical inference falls apart. In other studies, including Park, Marron and Rondonotti (2004) and Park, et al (2004b), it has been shown that such large estimates are often driven by unusual artifacts in the data, such as unexpected spikes with magnitude much larger than expected from conventional models, which is consistent with non-stationarity. This issue is discussed more deeply in Section 4.

Remark. A notion closely related to long range dependence is that of self-similarity (scaling). Strict self-similarity, typically associated with continuous time series $X(t)$, $t \in \mathbb{R}$, means that the series looks statistically similar over different time scales. Formally, there is a constant $H > 0$ such that the series $X(ct)$ and $c^H X(t)$ have the same finite-dimensional distributions for any $c > 0$. Self similar stochastic processes are Long Range Dependent, but LRD processes need not be self-similar, although LRD processes which show self-similarity over a substantial range of time scales appear frequently in network traffic. For more information, see the LRD references indicated above.

An important example of a LRD stochastic process is Fractional Gaussian Noise (FGN), ϵ_t, see for example Mandelbrot and Van Ness (1968), and Taqqu, et. al. (1995) for example. It is the increment of Fractional Brownian Motion $B_H(t), t \in \mathbb{R}$, i.e., $\epsilon_t = B_H(t) - B_H(t-1), t \in \mathbb{Z}$. $B_H(t)$ is the only Gaussian self-similar process with stationary increments. It satisfies $B_H(0) = 0$, has mean 0 and covariance

$$E\left(B_H(t)B_H(s)\right) = \frac{C}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right),$$

where $0 < H < 1$ and $C = \mathrm{Var}B_H(1)$. It follows that FGN is a stationary Gaussian time series with mean zero and autocovariance function $\gamma(t) = (|t+1|^{2H} + |t-1|^{2H} - 2|t|^{2H})/2$, $t \in \mathbb{Z}$. In the case that $H = 1/2$, it follows that $\gamma(t) = 0$ for $t \neq 0$, and so ϵ_t is an independent time series. For $H \neq 1/2$, as $t \to \infty$, $\gamma(t) \sim H(2H-1)t^{2H-2}$ and we have long range dependence for $1/2 < H < 1$. FGN is one of the simplest examples of long range dependent time series. A simple model for the time series of bin counts X_t is $X_t = \mu + \epsilon_t$, which has only three parameters, μ, $\sigma^2 = \mathrm{Var}(\epsilon_t)$, and Hurst parameter H. In the Dependent SiZer approach to analyzing trends in time series data, discussed and used in Section 4.1, this model is used as the null hypothesis.

Three methods have been used to simulate the FGN in this paper. The wavelet method to simulate fractional Brownian motion is based on a fast (bi-) orthogonal wavelet transform. More specifically, the method starts with a discrete time FARIMA$(0, H+1/2, 0)$ series of a relatively short length and produces another much longer FARIMA series that at the end, is subsampled to get a good approximation to fractional Brownian motion. For more details, see Abry and Sellan (1996) and Pipiras (2003). The spectral method is based on the fact that the periodic repetition of the covariance function of FGN on the interval [0,1] is a valid covariance function, and the covariance matrix of a time series with periodic covariance function is a circulant matrix. The Fast Fourier Transform (FFT) is used to diagonalize a circulant matrix, which thus gives fast and exact simulations of the FGN. For more details, see Wood and Chan (1994) and Dietrich and Newsam (1997), or the monograph Chilès and Delfiner (1999). The Fourier transform method is based on synthesizing sample paths that have the same power spectrum as FGN. These sample paths can be used in simulations as traces of LRD network traffic. For more details, see Paxson (1995) or Danzig, et al (2000).

For the simulated traces studied below, we generated time series of generally the same length as the UNC Link Data. We took the Hurst Parameter to be $H = 0.9$, which is consistent with the estimates that we obtain. We took the variance to be $\sigma^2 = 20$, motivated by the facts that the Poisson variance is the same as the mean and that the UNC Main Link generally had around 20 packets per millisecond. The deeper analysis in Park, Marron and Rondonotti (2004) revealed that a somewhat larger variance might have been more appropriate.

2.1 Hurst parameter estimation methods

Three major Hurst parameter estimation methods are considered here. The Aggregated Variance method is described in Section 2.1.1. Section 2.1.2 gives a definition of the Local Whittle approach. Introduction to the Wavelet Method can be found in Section 2.1.3. In addition to the short summaries provided here, more detailed information is available from links at the top of the web page Le, Hernández-Campos and Park (2004).

2.1.1 Aggregated Variance

A LRD stationary time series of length N with finite variance is characterized by a sample mean variance of order N^{2H-2} (Beran (1994)). This suggests the following method of estimation.

1. For an integer m between 2 and $N/2$, divide the series into blocks of length m and compute the sample average over each k-th block.

$$\bar{X}_k^{(m)} := \frac{1}{m} \sum_{t=(k-1)m+1}^{km} X_t, \qquad\qquad k = 1, 2, \ldots, [N/m],$$

2. For each m, compute the sample variance of $\bar{X}_k^{(m)}$ across the blocks

$$s_m^2 := \frac{1}{([N/m]-1)} \sum_{k=1}^{[N/m]} (\bar{X}_k^{(m)} - \bar{X})^2$$

3. Plot $\log s_m^2$ against $\log m$.

For sufficiently large values of m, the points should be scattered around a straight line with slope $2H-2$. In the case of short range dependence ($H = 0.5$), the slope is equal to -1. It is often convenient to draw such a line as reference, however, small departures from short range dependence are not easy to detect visually.

The estimate of H is the slope of the least squares line fit to the points of the plot. In practice, neither the left nor the right end points should be used in the estimation. On the left end, the low number of observations in each block introduces bias due to short range effects. On the right end, the low value of $[N/m]$ makes the estimate of s_m^2 unstable. The two thresholds that control the critical estimation range are typically left to the discretion of the researcher.

One disadvantage of the AV method is that it does not provide an explicit estimation of the variance of the estimator. Another disadvantage is that it has been found not to be very robust to departures from standard Gaussian assumptions (Taqqu and Teverovsky (1998)).

We used the `plotvar` function for Splus, see Sherman, Willinger and Teverovsky (2000).

In our analysis, we considered two choices for the thresholds for the linear fit. The "automatic choice" was $(10^{0.7}, 10^{2.5})$ in all cases. We also considered a careful tuning, where this choice was made individually for each data set, by manually trying to balance the trade-off that is expected, in particular by choosing a stretch that looked linear.

2.1.2 Local Whittle

The Local Whittle Estimator (LWE) is a semiparametric Hurst parameter estimator based on the periodogram. It was initially suggested by Kunsch (1987) and later developed by Robinson (1995). It assumes that the spectral density $f(\lambda)$ of the process can be approximated by the function

$$f_{c,H}(\lambda) = c\lambda^{1-2H} \tag{4}$$

for frequencies λ in a neighborhood of the origin.

The periodogram of a time series $\{X_t,\ 1 \leq t \leq N\}$ is defined by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} X_t e^{i\lambda t}\right|^2$$

where $i = \sqrt{-1}$. Usually, it is evaluated at the Fourier Frequencies $\lambda_{j,N} = \frac{2\pi j}{N}$, $0 \leq j \leq [N/2]$.

The LWE of the Hurst parameter, $\widehat{H}_{LWE}(m)$, is implicitly defined by minimizing

$$\sum_{j=1}^{m} \log f_{c,H}(\lambda_{j,N}) + \frac{I_N(\lambda_{j,N})}{f_{c,H}(\lambda_{j,N})} \tag{5}$$

with respect to c and H, with $f_{c,H}$ defined in (4).

The LWE depends on the number of frequencies m over which the summation is performed. It should be chosen so as to balance the trade-off between adding more bias as m increases, due to the fact that the approximation (4) holds only in a neighborhood of 0, and increasing the variance as m decreases. Henry and Robinson (1996) provided a heuristic computation of the bias and variance for this estimator.

The LWE has been proved to be fairly robust to deviation from standard assumptions (Taqqu and Teverovsky (1998)).

For computational convenience when employing the Fast Fourier Transform (FFT) algorithm to compute the periodogram, only 7257600 ($= 2^9 * 3^4 * 5^2 * 7^1$) observations per each data sets have been used. The periodogram has been calculated for the first 5000 Fourier frequencies and the optimization (5) performed by a simple quasi-Newton method for $m = 10, 20, ..., 5000$. Since (5) is effectively the negative log likelihood based on a Whittle approximation, we may also estimate standard errors from the observed information matrix. For each value of m, we can therefore calculate both a point estimate of H and an approximate 95% confidence interval, assuming the model (4) is exact for $0 < \lambda \leq \lambda_{m,N}$.

From inspection of many such curves (some of these are shown in Section 4.2), we have observed that in many of the present cases the values of H stabilize within the range $m \in (1000, 2000)$ and therefore suggest that $m = 2000$ is a reasonable overall value to take. In the following discussion we will call this the "automatic method". However, it raises the question of whether a data-based adaptive method would be superior.

Henry and Robinson (1996) proposed the following approach to choose m. The approximation is based on the assumption that the true spectral density f satisfies

$$f(\lambda) = c\lambda^{1-2H}(1 + E\lambda^{\beta} + o(\lambda^{\beta})), \quad \lambda \downarrow 0, \tag{6}$$

for some E and $0 < \beta \leq 2$. Based on (6) they derived the following approximation to the mean squared error (MSE):

$$\frac{1}{4}\left\{\frac{1}{m} + E^2 \frac{\beta^2}{(\beta+1)^4}\left(\frac{2\pi m}{N}\right)^{2\beta}\right\}. \tag{7}$$

Based on (7) one can easily solve for the optimal m. In practice they suggested fixing $\beta = 2$, estimating c and H by the local Whittle method for fixed m, and estimating E by simple regression of $I_N(\lambda_{j,N})/f_{c,H}(\lambda_{j,N})$ against $\lambda_{j,N}^2$ for $j = 1, ..., m$. The expression (7) is then optimized to define a new value of m and the whole process repeated.

For the data sets considered in this paper, this procedure was tried but in most cases did not converge. As an alternative, we have evaluated (7) for each of $m = 10, 20, ..., 5000$, using the values of c, H and E estimated from the first m periodogram ordinates, and chosen the value of m by direct search over the values of (7). We call this the "tuned" method. In nearly all cases this procedure led to a clear-cut choice for the best value of m.

2.1.3 Wavelet method

Abry and Veitch (1998) and Veitch and Abry (1999) proposed a wavelet based estimator of the Hurst parameter, H. For a quick introduction to wavelets (for simplicity, done in the continuous time domain), let ψ be a wavelet function with the moment order L, that is,

$$\int u^l \psi(u) du = 0, \quad l = 0, 1, \ldots, L-1, \tag{8}$$

(e.g., see Daubechies (1992)) and $\psi_{j,k}(u) = 2^{-j/2}\psi(2^{-j}u - k)$, $k \in \mathbb{Z}$. The wavelet coefficients of X_t, denoted $d_{X_t}(j,k)$ are defined as

$$d_{X_t}(j,k) = \int X_t(u)\psi_{j,k}(u)du.$$

If we perform a time average of the $|d_{X_t}(j,k)|^2$ at a given scale, that is,

$$\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} d^2_{X_t}(j,k),$$

where n_j is the number of coefficients at octave j, then

$$E\log_2(\mu_j) \sim (2H-1)j + C, \tag{9}$$

where C depends only on H. Define the variable Y_j for $j = j_1, \ldots, j_2$ as

$$Y_j \equiv \log_2(\mu_j) - g_j,$$

where $g_j = \Psi(n_j/2)/\ln 2 - \log_2(n_j/2), \Psi(z) = \Gamma'(z)/\Gamma(z)$, and $\Gamma(\cdot)$ is the Gamma function. Then, using the relationship (9), the slope of an appropriate weighted linear regression, applied to (j, Y_j) for $j = j_1, \ldots, j_2$, provides an estimate of the Hurst parameter. Discrete wavelet theory is entirely parallel to the continuous theory, which was considered here for simplicity of notation.

One of the advantages of applying the wavelet method to the Internet traffic data is that the vanishing moments, up to order L as defined at (8), allow the method to ignore polynomial trends in the data up to degree $L-1$. Thus this Hurst parameter estimation is much more robust against simple trends, than other methods considered here. Another advantage of the wavelet-based method is that for a stationary process X_t, the wavelet coefficients tend to be approximately uncorrelated, see Kim and Tewfik (1992), Bardet, et al (2000) and Bardet (2002).

Matlab code for implementing the wavelet method is available at the web site Veitch (2002) . There are two parameters to be selected in implementation, j_1 and L (with j_2 as large as possible at a given octave j). As for the Aggregated Variance Estimate, and the Local Whittle Estimate, we consider both "automatic" and "tuned" versions of this H estimator. For the automatic case, L is fixed to 3 and automatically j_1 is chosen according to the method of

Veitch, Abry, and Taqqu (2003). For the tuned case, several values of L and j_1 are investigated for each data set, and the optimal values are selected by considering the goodness-of-fit measure in Veitch, Abry, and Taqqu (2003) and the stability of the H estimates.

3 Summary of Main Results:

This section provides a summary of our analysis, which is detailed on the web page Le, Hernández-Campos and Park (2004). There are a number of comparisons of interest. In Section 3.1 we compare across Hurst parameter estimators, i.e. we compare the different methods with each other. In Section 3.2 we compare variations within each type of Hurst parameter estimator, including comparison of automatic with tuned versions, and study of trend issues. In Section 3.3 we investigate aggregation issues. An overview of these results is given in Section 3.4.

The main body of our web page, Le, Hernández-Campos and Park (2004), consists of tables with summaries of our results. Different tables represent different types of data. All tables allow downloading of the raw data (time series of packet counts, and some cases byte counts as well, for consecutive 1 ms time intervals), from links in the left hand column. The text in most of the columns, summarizes the estimated value of H, and (except for the Aggregated Variance method) gives an indication of the sampling variation. Each summary is also a link to one or more diagnostic plots for each case. The right hand column of each table provides a link to a SiZer analysis for each raw data time series, as discussed in Section 3.2.2.

3.1 Comparison across Estimators

This section summarizes results of our study of Hurst parameter estimation across the different methods. This comparison is between columns of the web page Le, Hernández-Campos and Park (2004).

3.1.1 Lab Results

Here Hurst parameter analysis is done for the lab generated data sets. These appear in the web page, Le, Hernández-Campos and Park (2004), in the top 5 rows of the first three tables. In this section we focus on just the top table. The estimated values of H are summarized as a parallel coordinate plot, see Inselberg (1985), in Figure 2. The 3 curves indicate the 3 different estimation methods. The horizontal axis indexes the particular lab experiment. An indication of the perceived sampling variation (computed differently for the different estimators), is given by the error bars. The perhaps too wide range shown on the vertical axis, allows direct comparison with the results in the next section.

A clear lesson from Figure 2 is that all of the different estimation methods give very similar results (differences are comfortably within the range suggested by the error bars). In particular, all estimated H values are approximately in the range $[0.82, 0.90]$. The consistency across estimators suggests that all of them are working rather well. The consistency across lab settings suggests that in this study, the Hurst parameter is very stable, even across this range of different traffic loads. By construction, the laboratory experiments produce stationary time series of packet counts. The emulated user population is held constant during an entire experiment as are the distributions used by the generating programs to select random file sizes, "think" times, etc. We note also that the results showing the presence of quite strong long range dependence are consistent with the theory developed by Willinger, Taqqu, Sherman and Wilson (1997). Even at the lowest load (2% or 20 Mega (million) bits per second (Mbps)), the aggregation of 4,000 ON/OFF sources (users) where ON times (response transfers) or OFF times (thinking) have heavy-tailed distributions appears to be sufficient for long range dependence (and a version of self-similarity that holds across a range of scales, see Section 3.3) in the packet counts time series. Aggregation of many more users (up to about 28,000) appears to have no direct effect on the Hurst parameter.

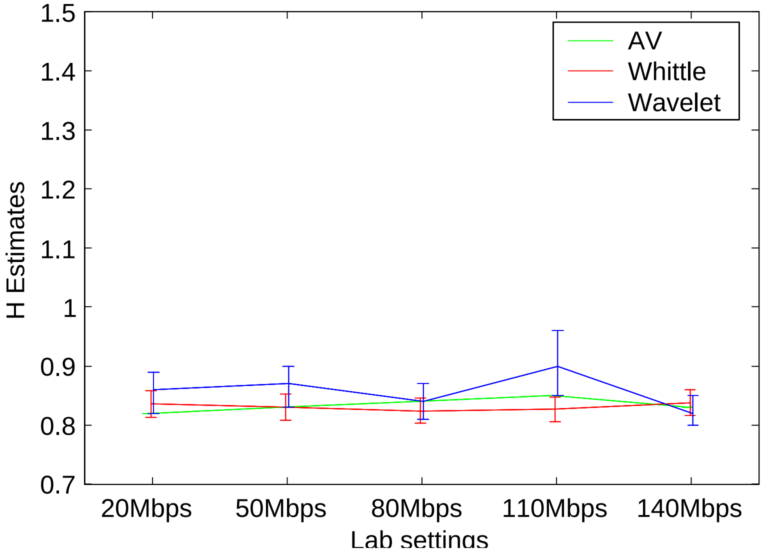

FIGURE 2: *Summary of estimated Hurst parameters, for Lab data. Curves are different estimation methods. Different lab settings on the different axes. Error bars represent perceived likely variation in estimates. Shows very consistent estimation for lab data.*

A major motivation for this study is the comparison of lab generated data with real data. This is done in the next section.

3.1.2 Link Results

In this section, the parallel analysis to Figure 2 is done for the UNC Main Link Data. These appear in the web page, Le, Hernández-Campos and Park (2004), in the remaining rows of the first three tables. Again the focus is on just the top table.

The estimated Hurst parameters are summarized in Figure 3, again as a parallel coordinate plot, with thick curves showing the different estimators. This time the horizontal axis represents different time periods. The time ordering (as in the web page) was not particularly insightful, so instead they are grouped according to time of the day.

In Figure 3, the sampling variation is shown using appropriately colored dotted lines upper and lower limits, because there are too many time periods to represent these as intervals as in Figure 2.

Figure 3 shows a much larger range of variation of estimated Hurst parameters, in particular over $H \in [0.79, 1.48]$. The variation happens over *both* type of estimator, and also time period. The very large differences in estimator type suggest that the Hurst parameter may not be the most sensible way to quantify the structure in these time series. In addition, a number of estimates $H > 1$, raise serious issues about the validity of the underlying models. Simulation of processes with this structure can not be very well done using straightforward models of the type that are simply parametrized by the Hurst parameter. This is in stark contrast to the lab data, analyzed in Figure 2, which shows that the lab data are quite different from the UNC Link Data, and which motivates the development of richer models for simulation.

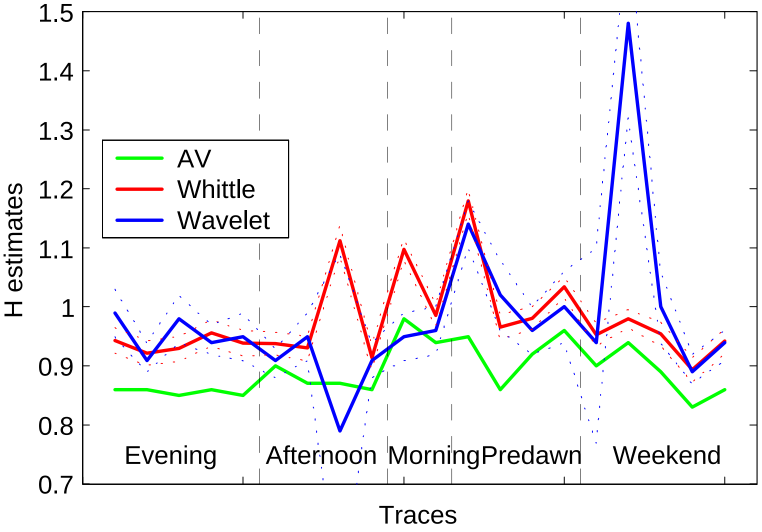

FIGURE 3: *Summary of estimated Hurst parameters, for UNC Link data. Thick curves are different estimation methods. Different time blocks appear on the axes. Dotted curves reflect perceived likely variation in estimates. Shows very divergent estimation for different methods, and also for different time blocks.*

There are systematic differences across estimator type. The Aggregated Variance method never has $H > 1$, because of a constraint in its definition, and is typically smaller than the other methods, except for Thursday Afternoon. There are a number time periods where the Whittle method is larger than the others, perhaps most noticeably on Tuesday Morning and Thursday Afternoon. Some of this is due to trend issues, which are discussed in Section 3.2.2. The Wavelet Hurst parameter estimates are generally quite stable, and lie between the other estimates. Unusual behavior of the wavelet estimates occurs for Thursday Afternoon, Tuesday Predawn (after midnight, before 6:00AM) and Saturday Afternoon. It will be seen in Section 4 that this is due to serious nonstationarity in these data sets.

There is also systematic structure within time periods. All Hurst parameter estimates are effective and consistent with each other for the evening time periods. This is roughly true for afternoons as well, with the exception of Thursday Afternoon. The morning, predawn time and weekend time periods have far less stability and consistency across estimators.

In Section 3.2.2 it is seen that one cause of the morning-predawn-weekend instability is major trends during these time periods. But deeper investigation has revealed much more complicated types of non-stationarity, as discussed in Section 4, where the unusual cases are explicitly studied.

A different type of visual summary, using scatterplots of Hurst parameter estimates, is available from the *Comparative Plots of Hurst Parameter Estimation Methods* link on the web page Le, Hernández-Campos and Park (2004). This is not shown here because results are essentially the same as shown in Figures 2 and 3, and to save space.

3.1.3 Simulated results

Parallel results for several data sets simulated from Fractional Gaussian Noise processes, with true Hurst parameter $H = 0.9$ (using different FGN simulation algorithms), appear in the lower two tables. Rows of the table refer to the different simulation methods discussed in Section 2. The data in the first row of the table was generated using the spectral method, the second using the wavelet method, and the third using the Fourier approach.

Again the results from the first of these tables is summarized using a parallel coordinate plot in Figure 4. As for Figure 2, the vertical axis in Figure 4 is taken to be the same as that for Figure 3 for easy comparison (even though this is poor use of plot area). The Hurst parameter estimation results for the simulated data are even more steady than for the lab data, shown in Figure 2. In particular, $H \sim 0.86 - 0.9$ across all estimation methods and all simulation algorithms (and these are much closer to 0.9 if the Aggregated Variance is ignored). An important conclusion is that all estimation methods perform extremely well in the setting for which they were originally designed.

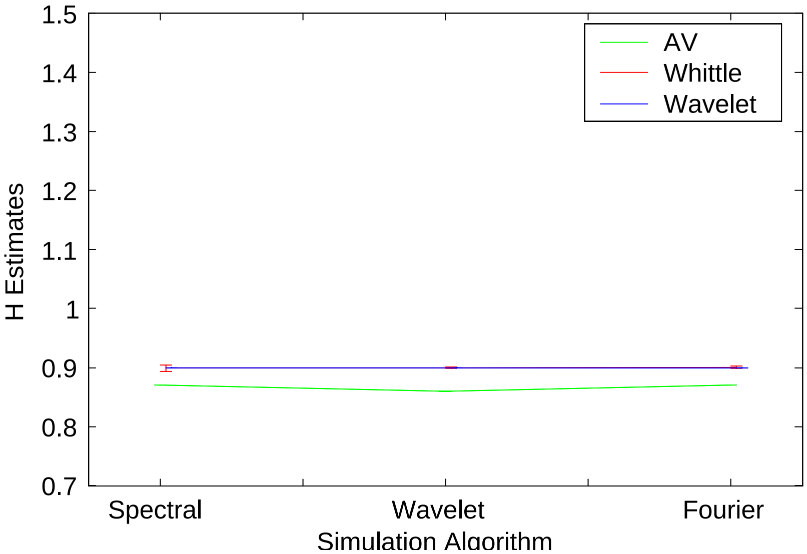

Figure 4: *Summary of estimated Hurst parameters, for simulated FGN data. Curves are different estimation methods. Different simulation algorithms on the different axes. Error bars represent perceived likely variation in estimates. Shows very consistent estimation.*

The major conclusion of this section is that, from the Hurst parameter point of view, the lab data are closer to simulated FGN data, than to the real UNC Main Link data.

3.2 Comparison of Variations Within estimators

In this section we compare variations of each of the Hurst parameter estimators. Manual tuning is considered in Section 3.2.1. The impact of simple trends are considered in Section 3.2.2.

3.2.1 Tuning

As noted in Section 2.1, all of the Hurst parameter estimation methods considered here allow tuning in various ways. In Section 3.1, we compared the approaches to Hurst parameter estimation using recommended defaults. Here we study how those defaults compare to more careful tuning.

Figure 5 does this comparison using a scatterplot. Each dot represents a row of the data tables on the web page, Le, Hernández-Campos and Park (2004). the numbers above each dot are the row numbers, so dots 1-5 correspond to lab data traces, and 6-25 indicate the UNC Link data traces. The horizontal coordinate of each dot is the manually tuned estimate of H, while the vertical coordinate is the automatic version, studied in Figures 2 and 3 above. The blue line is the 45 degree line, included for reference as to where the coordinates are the same.

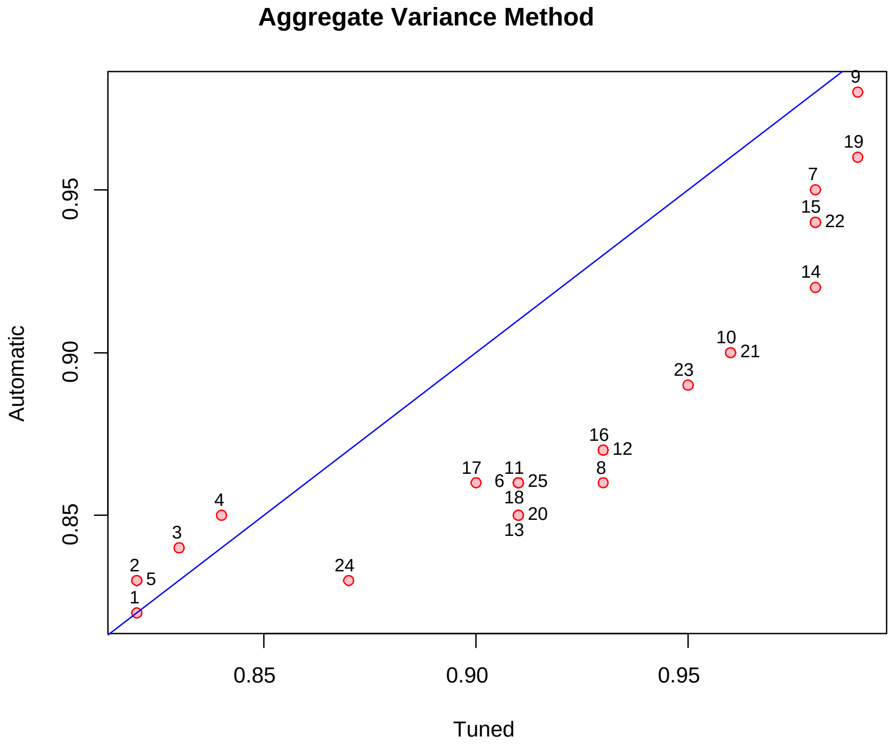

FIGURE 5: *Scatterplot summary of automatic versus manually tuned versions of Aggregated Variance estimated Hurst parameters, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows systematic bias between manual and automatic versions.*

The lab data, 1-5, appears in the lower left, which is consistent with the green curve for the lab data in Figure 2 taking on generally lower values than the green curve for the UNC Link data in Figure 3. These dots are also close to the 45 degree line, indicating that the manually tuned H estimates are quite close to the automatic default choices. For the UNC Link data, 6-25, the dots lie well below the 45 degree line, and suggest a strong systematic bias in the direction of larger H estimates for the manually chosen versions. This can be interpreted in several ways, but inspection of the diagnostic plots suggests that in general the Aggregated Variance defaults are not effective for this type of data. This is consistent with the fact that the Aggregated Variance estimates seem generally smaller than the others. We conclude that an important weakness of the Aggregated Variance method is this strong need for manual intervention.

Similar investigation is done, for the Local Whittle approach to H estimation, in Figure 6. Again the dots correspond to data sets, with the same numbering scheme. The horizontal and vertical coordinates are again tuned and automatic choices of the range for linear fitting.

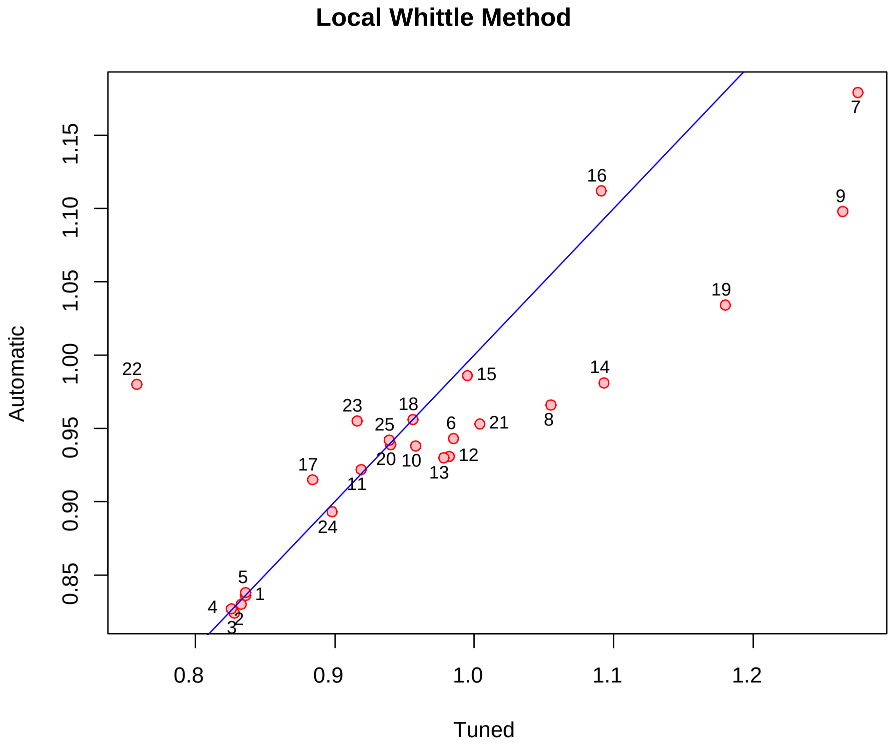

Figure 6: *Scatterplot summary of automatic versus manually tuned versions of Local Whittle estimated Hurst parameters, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows stronger consistency between automatic and tuned versions.*

Fig. 6 shows a plot of automatic against tuned Whittle estimates of H. Again points 1–5 are for the five lab data sets and 6–25 are for the 20 real data sets. With the sole exception of data series 22 - Saturday Afternoon, the plot shows that the tuned estimates of H are in general similar to or larger than the automatic values. Inspection of the detailed plots (some of these are shown in Section 4.2) shows that in most cases where the tuned estimate of H is much larger than the automatic value, with a lot of instability in H, and the tuned value of m is generally much less than the default 2000 used for the automatic estimate. Several of these series are also cases where later analysis shows the presence of trends, as discussed in Section 3.2.2.

Figure 7 shows a similar scatterplot for comparing the automatic and tuned Hurst parameter estimates for the Wavelet approach. The format and labelling are the same as for Figures 5 and 6.

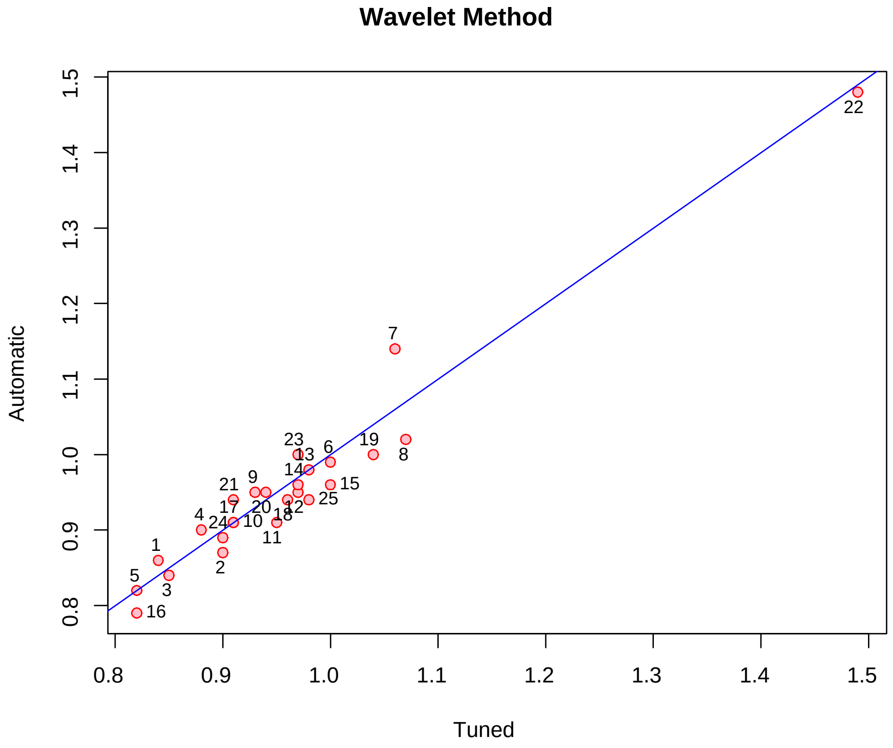

FIGURE 7: *Scatterplot summary of automatic versus manually tuned versions of Wavelet estimated Hurst parameters, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows very strong consistency between automatic and tuned versions.*

The main lesson of Figure 7 is that there is little difference between the automatic and tuned choices, because all of the dots lie close to the blue 45 degree line. The largest departure from this is data set 7 - Tuesday Predawn, which was also noted as giving unusual wavelet behavior in Section 3.1.2. Data set 22 - Saturday Afternoon is very noticeable as having an unusually large Wavelet Hurst parameter estimate, although the automatic and tuned choice are quite similar. These unusual behaviors will be studied more deeply in Section 4.

3.2.2 Detrending

An important underlying assumption of all Hurst parameter estimators is that the data come from a stationary stochastic process. This will typically not be true for Internet traffic data, because of diurnal effects. For example, at a university, there will typically be many more users during the business day, than during off hours. In the two hour time block studied here, these effects will usually appear as approximately linear trends in some of the time periods. We study such trends in this section, using a SiZer analysis of the trends in the 9 - Tuesday Morning trace, as shown in Figure 8. SiZer analyses for all of the traces studied here are available from the links in the right column of the web page Le, Hernández-Campos and Park (2004).

The top panel of Figure 8 shows a sampling of the bin counts as green dots (hence the vertical coordinates are integers). The blue curves are a family of moving window smooths, for a wide range of different window widths (a multiscale view), which thus range from quite smooth to rather wiggly. But there is a clear visual upward trend to all of the blue curves. Is the trend statistically significant? Are the wiggles significant?

SiZer (SIgnificance of ZERo crossings) was proposed by Chaudhuri and Marron (1999), to address this type of question. The bottom panel of Figure 8 shows a SiZer map. The SiZer map is an array of colored pixels. The horizontal locations correspond to the times (the horizontal locations in the family of smooths in top panel). The vertical locations correspond to the level of smoothing (essentially the width of the smoothing window), i.e. the scale. In particular, each row of the SiZer map does statistical inference for one of the blue curves in the top panel. The inference focuses on the slope of the curve. When the slope is significantly positive, the pixel is colored blue. Red is used to indicate that the slope is significantly negative. When the slope is neither (i.e. the estimated slope is close enough to 0 that there is not significant difference), the intermediate color of purple is used.

The SiZer map for Figure 8 shows exclusively blue for all of the larger windows (coarser scales), i.e. the upper rows in the SiZer map. This is evidence of a strong upwards trend. Substantial red regions appear in the lower part of the map, i.e. the finer levels of resolution or smaller window widths. These correspond to short term decreases in the more wiggly blue curves, which appear as wiggles in the top panel. While the wiggles may appear to be spurious, the red coloring shows that these are statistically significant (with respect to a null hypothesis of white noise). One reason for this significance is that the very large sample size of 7,200,000 allows enormous statistical power. Data sets of the same size generated as FGN will exhibit similar structure, because these paths have wiggles that are larger than expected under the white noise null assumption. This suggests that a useful modification of SiZer would allow substitution of FGN for the white noise as the null hypothesis. This idea has been implemented by Park, Marron and Rondonotti (2004), and that tool is used in Section 4 below.

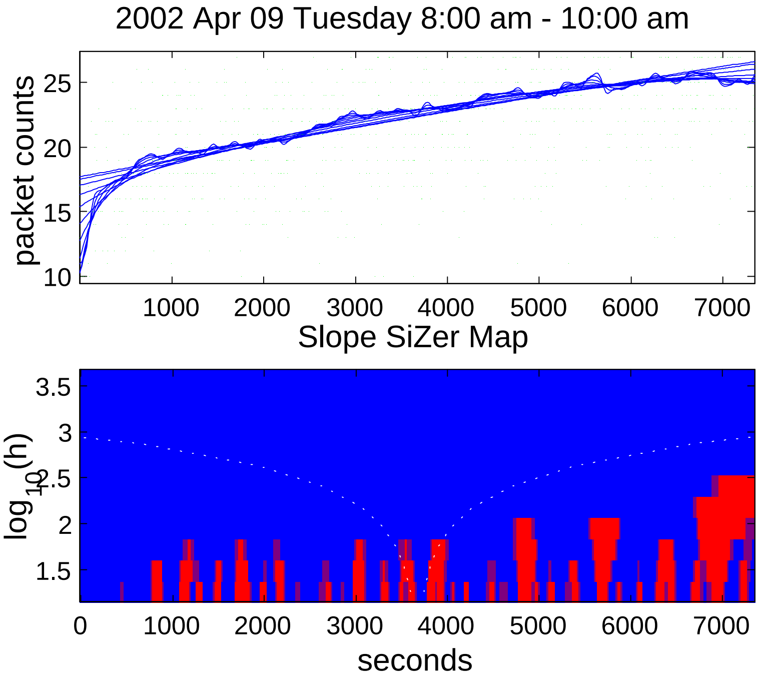

Figure 8: *SiZer analysis of trends in the Tuesday Morning trace. Shows strong increasing trend, that is well approximated by a line.*

Figure 8 shows a clear increasing trend in the data, which violates the assumption of stationarity, which is fundamental to even the definition of the Hurst parameter. As noted above, these trends appear to be due to diurnal effects, and on the time scales considered here, are reasonably well approximated by linear trends. Hence, a simple and direct way to study the impact of these trends on the estimation process is to remove the linear trend from each data set (essentially subtract the least squares linear fit from the data). This type of detrended data appears in the third table of the web page Le, Hernández-Campos and Park (2004). As for other data sets, the detrended data may all be downloaded from the links on the left, and the analogous Hurst parameter estimates are computed, using automatic versions of each.

Figure 9 shows a scatterplot, in a similar format, with the same numbers (indicating rows in the results table).

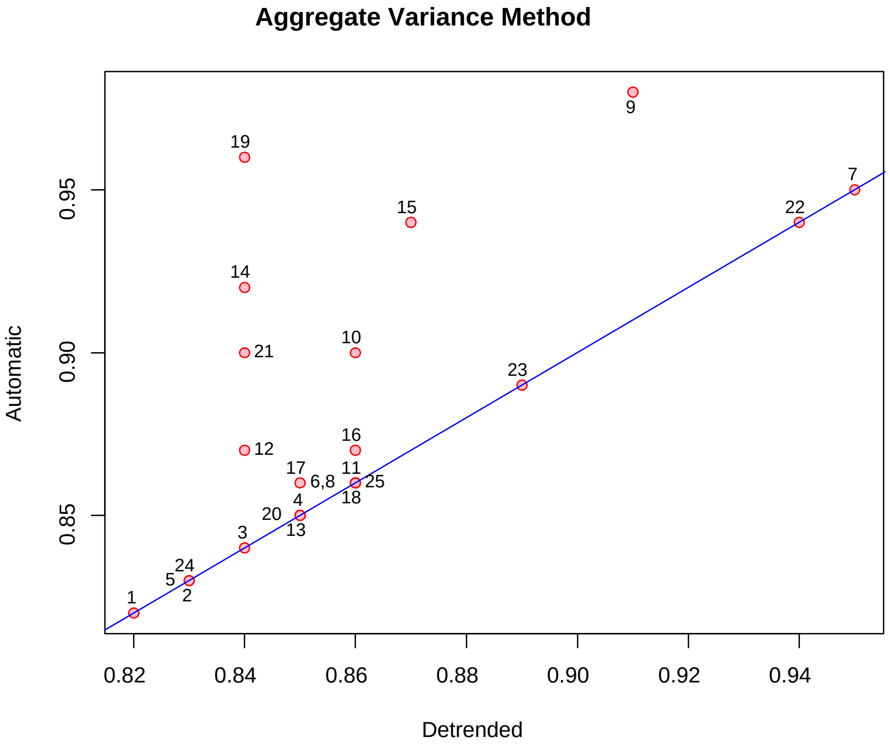

FIGURE 9: *Scatterplot summary of raw versus detrended Aggregated Variance Hurst parameter estimates, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows that Aggregated Variance is strongly affected by linear trends.*

While some data sets in Figure 9 lie right on the 45 degree line indicating no difference between whether the trend is removed or not, others show a very marked difference. An inspection of the SiZer maps for the data sets near the 45 degree line, including all of the lab data sets (which are designed to have no trend) and some of the UNC Link data sets, shows no significant (or else very small) linear trend . When there is a difference in the H estimates, the detrended data always result in a smaller estimated H, suggesting that linear trends tend to increase the Aggregated Variance estimate. The cases with the largest difference are 9 - Tuesday Morning, 10 - Tuesday Afternoon, 12 - Wednesday Afternoon, 14 - Thursday Predawn, 15 - Thursday Morning, 19 - Friday Predawn, and 21 - Saturday Morning. An inspection of the corresponding SiZer maps shows that all of these are time periods of strong trend, that is well approximated by a line.

We conclude that linear detrending is essential before using the Aggregated Variance method of Hurst parameter estimation.

Figure 10 shows a similar comparison, between the raw and detrended data, for the Local Whittle Estimate.

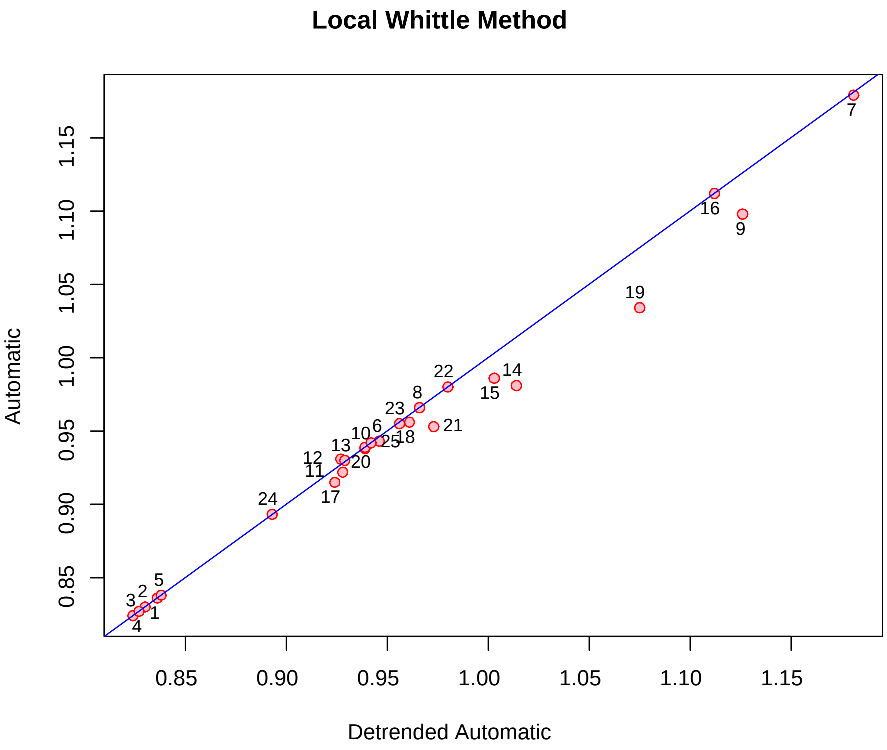

FIGURE 10: *Scatterplot summary of raw versus detrended Local Whittle Hurst parameter estimates, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows that the Local Whittle estimate is much less affected by linear trends.*

The points in Figure 10 are much closer to the diagonal for the Local Whittle method, than they were for the Aggregated Variance, as shown in Figure 9. This suggests that the Local Whittle method is less affected by linear trends in the data. Where there is any difference, the H estimate is larger for the detrended data than for the raw data. These are the cases in which there appears to be a significant trend, as noted above.

Similar comparison between raw and detrended data, based on the Wavelet Hurst parameter estimate, is shown in Figure 11.

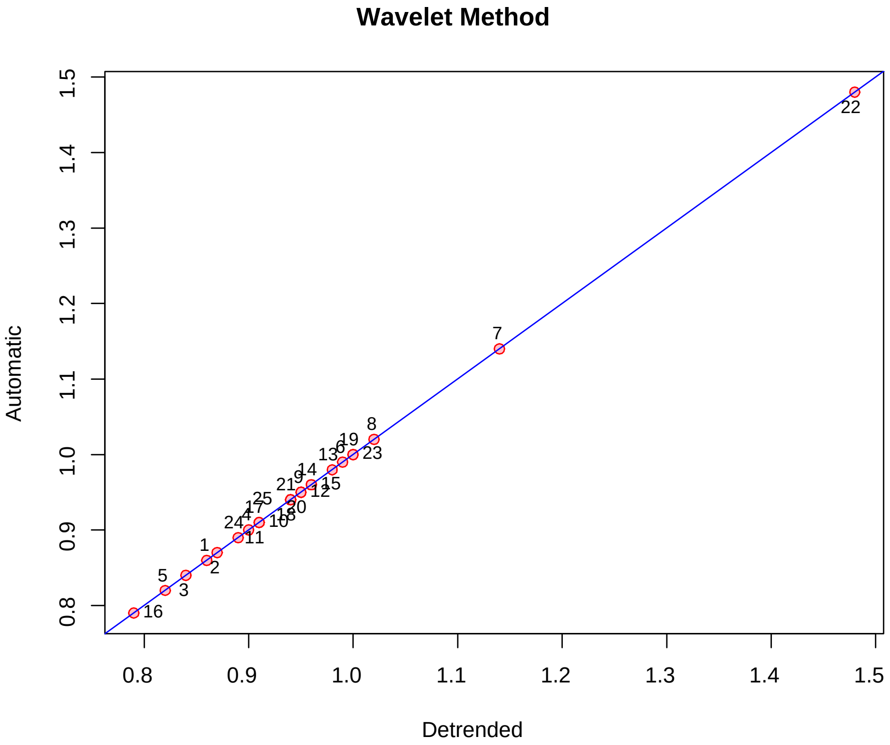

FIGURE 11: *Scatterplot summary of raw versus detrended Wavelet Hurst parameter estimates, over Lab and UNC Link data sets. Numbers are rows of table in web page. Shows that the Wavelet estimate is completely unaffected by linear trends.*

For the wavelet estimator, all dots lie right on the 45 degree line, which confirms the expectation (since the wavelet approach essentially does automatic linear, and even quadratic, since $L = 3$ in (8), trend removal) that the Hurst parameter estimate is the same for both raw and detrended data. As it was designed to be, the wavelet estimator is very robust against linear trends.

While adjusting for linear trends explains some of the unusually large Hurst parameter estimates, there are still many others that are present. It will be seen in Section 4, that these tend to be caused by other types of nonstationarity. An example of nonlinear trend can be seen in Figure 8, where the trend is mostly linear (constant slope), but is much steeper near the left edge. This type of trend affects even the wavelet Hurst parameter estimate, since it is not effectively modelled by a low degree polynomial.

3.3 Aggregation

All data sets analyzed here are time series of bin counts with 1 millisecond bins. An interesting question is how the results differ at other levels of aggregation. The self-similar scaling ideas suggest that the analysis should be unchanged, but such ideas do not always hold up for UNC Link data, so we investigated a wide range of other aggregations. These can be seen in many of the diagnostic plots linked from the web page Le, Hernández-Campos and Park (2004).

A summary for the wavelet case appears in Figure 12. This shows explicit results for the automatic version of the Wavelet Hurst Parameter estimator case. The Manually Tuned Wavelet and the Local Whittle cases are not shown, because the main lessons are the same.

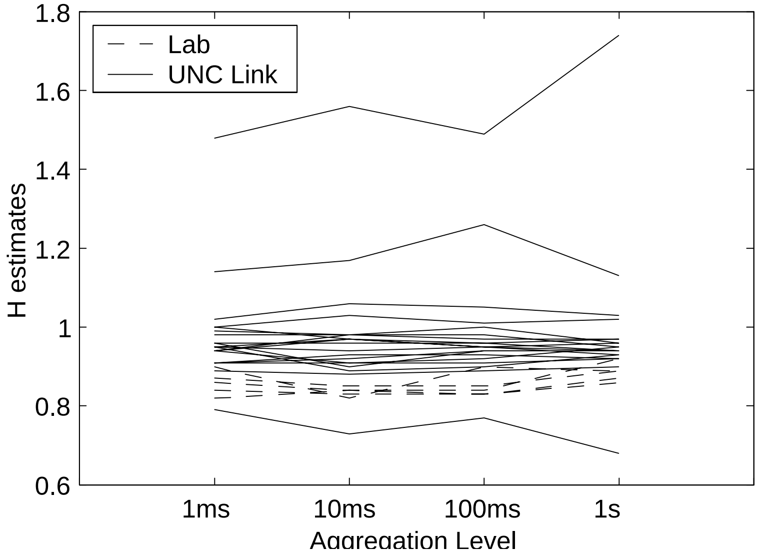

FIGURE 12: *Parallel coordinate plot of Wavelet Hurst parameter estimates for different levels of aggregation. Shows little change in estimates.*

Figure 12 shows the effects of aggregation, from the original binwidth of 1 ms, through 10 ms and 100 ms, all the way to a binwidth of 1 sec. Each curve corresponds to one data set, with the lab data shown as dashed curves, and the UNC Link data shown as solid ones. The changes in estimated Hurst parameters, over this wide range of scales, are quite small and do not indicate a drift in any particular direction. This suggests that the well known self-similar scaling property holds up very well over this range of scales, for these data.

Three of the cases might be considered exceptions to this general constancy, the two above and the one below the main bundle. These are 22 - Saturday Afternoon, 7 - Tuesday Predawn, and 16 - Thursday Afternoon. It is seen in Section 4 that each of these exhibits different types of strong nonstationarity. While these curves wiggle somewhat, we were actually surprised that self-similarity appears to hold this well in the face of the very strong non-stationarities present in these data sets.

3.4 Overview and Conclusions

The above analysis studies Hurst parameter estimation from a number of different viewpoints, for a variety of purposes. This allows comparison at a number of different levels. These include a wide range of both data types, and also a large array of data within types. Another important comparison is across H estimation methods, and among variations within methods. In addition, we have provided further confirmation of existing ideas. Finally, we have seen that stationarity issues are paramount, so this will be further discussed in Section 4.

A lesson which cut clearly across all of the different analyses above is that there are major differences between the simulated data, the lab data and the UNC Link data. The simulated data examples all gave excellent Hurst parameter estimation, showing that when all assumptions are precisely correct, then all H estimators perform very well. The lab data aims to more closely mimic the UNC Link data, and does a good job of replicating a number of hard to simulate properties, such as LRD, with a Hurst parameter similar to that often encountered in real data as shown in Figures 2 and 3, plus self-similar scaling properties as shown in Figure 12. However a second issue that cuts across all of the analyses is that from study of a large number of real data traces, we see that there are also very important differences between the lab data and a number of the UNC Link traces. Deeper investigation shows that this is due to various types of non-stationarities, which will be more deeply discussed in Section 4. This motivates some ideas for potential improvements in the simulation process, discussed in Section 4.4.

Sections 3.1 and 3.2 provide a detailed comparison across Hurst parameter estimation types. There was a distinct ordering of the methods, with Aggregated Variance typically smaller, the Local Whittle largest, and the Wavelet estimate of H in between. While all methods perform well for the simulated data, some were much more robust against the unfortunate violation of assumptions that is endemic to Internet traffic, and also quite different amounts of manual intervention were found to be important. The Aggregated Variance method was least robust, and fortunately was also the method requiring most manual tuning. The Local Whittle method was much better on both accounts. The Wavelet approach was somewhat better in terms of decreased need for tuning, and had the expected solid robustness properties, often due to its implicit filtering out of low degree polynomials. We generally recommend that the Wavelet method be the basis of future work on Hurst parameter estimation, at least in the context on Internet traffic. A further advantage of the Wavelet approach is that the wavelet spectrum is a useful diagnostic, as discussed in Stoev, et al (2004a). However, some researchers prefer the Fourier transform view of time, so we recommend the Local Whittle approach for them. For this approach, it is also important to study diagnostics, as illustrated in Section 4.2.

In general, our results support some important existing notions about the nature of Internet traffic. The now widely accepted notion of long range dependence is strongly confirmed by the fact that all estimated Hurst parameters considered here are much larger than the 1/2 which would indicate short range dependence. In fact typical estimated values of H are often 0.9, or larger, suggesting very strong LRD. Another commonly held view, that is confirmed by Figure 12, is the self-similar scaling property, which is seen to hold over a wide range of scales from 1 millisecond to 1 second.

While we have confirmed the value of a number of conventional models and methods, the same analysis shows common flagrant violation of the assumptions that underlie these. In particular, non-stationarity is a very serious issue. While conventional linear trend non-stationarities are seen to be a serious issue in Section 3.2.2, much more challenging types of non-stationarity are also quite common. These non-stationarities frequently generate Hurst parameter estimates that are much larger than 1. This is at least conceptually problematic, because it is a violation of assumptions which underlie the mathematics of the Hurst parameter estimation process itself. This important issue is discussed further in Section 4.

4 Non-stationarity

A major lesson of the detailed Hurst parameter estimation discussed in this paper is that stationarities are a vital issue. Such non-stationarities were previously found by Paxson and Floyd (1995), Roughan and Veitch (1999) and Cao et. al. (2001). Our focus has been on the UNC Link, where average traffic rate varies over time, with its maximum in a day about half again as large as the minimum. The variance of the traffic trace changes with time as well. As noted in Section 3.2.2, because the lengths of the traffic traces we studied are two hours, the average traffic rate may remain a constant, have an upward trend or have a downward trend, depending on the particular time window of the day the traffic trace is taken. In this section we focus on nonlinear types of stationarity. A fundamental tool is Dependent SiZer, discussed in the next section.

4.1 Dependent SiZer Analysis

Dependent SiZer was developed in Park, Marron and Rondonotti (2004), as a tool to find non-stationarities in LRD time series. It is a variation of the SiZer map, as shown in Figure 8. Recall that in Figure 8, at the finer scales, there were many wiggles in the smooth, which were flagged as statistically significant by the red and blue regions in the SiZer map. This was natural, because the null hypothesis of conventional SiZer is a white noise process, and LRD processes such as FGN are expected to generate far larger wiggles. Dependent SiZer changes this null hypothesis to a given FGN. Thus it aims to *find structures in the data, that are not explainable by FGN.* Details on the choice of the FGN parameters, H and σ^2, used here can be found in Park, Marron and Rondonotti (2004).

Dependent SiZer analysis of the data sets considered in this paper are in the web page Park (2004), which can be accessed from the link *Dependent SiZer Web page* of Le, Hernández-Campos and Park (2004). The results are summarized, for each major type of data, in the following sections.

4.1.1 Simulated results

For all of these, the SiZer maps are almost all purple as expected, because these data were simulated from the null hypothesis. These are not shown here to save space, but are available in the Simulated FGN Data block of the web page Park (2004). But this analysis is worthwhile, because it reveals an area for potential improvement of Dependent SiZer: all of these plots indicate statistically significant structure in the second finest scale (and not the finest!). Future work is intended to understand this problem, and also to sharpen the statistical inference, using ideas from Hannig and Marron (2004).

4.1.2 Lab Results

For the laboratory generated data, the SiZer maps tended to be purple, suggesting that all structures in data could be explainable as artifacts of FGN. This is consistent with our understanding of the mechanisms used to generate the data. We considered both null hypotheses of $H = 0.8$, and $H = 0.9$.

These choices were interesting because the estimated automatic Wavelet Hurst parameter estimates were $H \in [0.82, 0.9]$. Hence, it is not surprising that all SiZer maps were essentially completely purple for the null of $H = 0.9$. For the null of $H = 0.8$, the lighter traffic loads of 20, 50 and 80 Mbps gave all purple SiZer maps, but there were a substantial number of changes in slope that were flagged as statistically significant for the larger traffic loads of 110 and 140 Mbps, as shown in Figure 13.

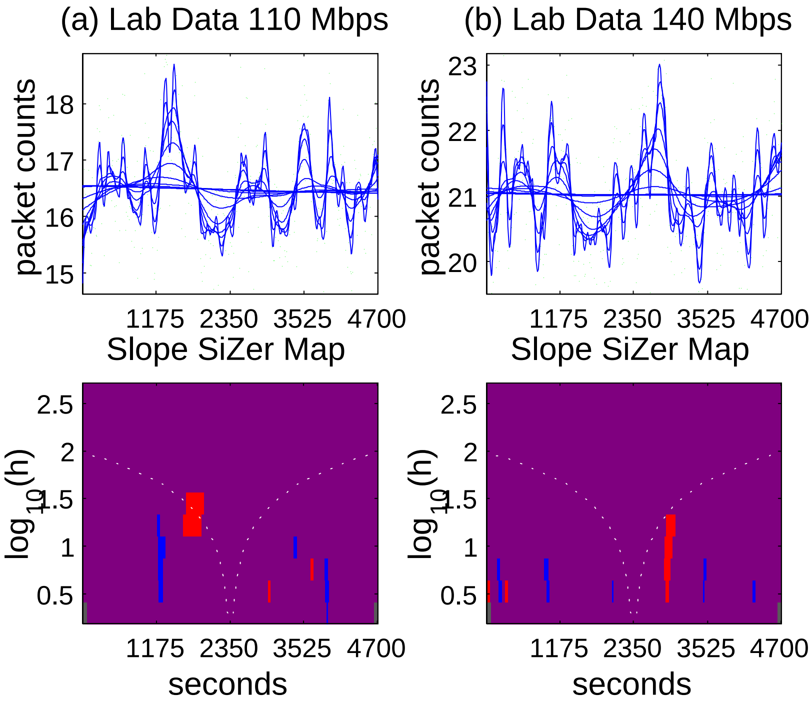

FIGURE 13: *Dependent SiZer analysis of the 110 Mbps and 240Mbps Lab data. Shows significant bursts under the Null Hypothesis of* $H = 0.8$*, for both settings.*

This increase in burstiness found by Dependent SiZer is very consistent with expected TCP effects. As the load increases, there is more packet loss, leading to an expectation of greater burstiness. Perhaps surprising is that the Hurst parameter estimates studied in Figure 2 do not show the same trend. This suggests that, contrary to the ideas of some, the Hurst parameter is at best a very weak surrogate as a "burstiness parameter".

4.1.3 Link Results

As noted in Section 3, a few of the UNC Link data time periods suggested strong non-stationarities in a number of ways. In this section, those data sets are studied more carefully. The Dependent SiZer analysis of all data sets, summarized on the web page Park (2004), shows that these were indeed the time periods with structure beyond what can be explained by standard FGN, with perhaps an additive linear trend. The fact that there are only 4 such suggests that such a model could be effective for simulation in many situations.

Data Set 7 - Tuesday Predawn Potential non-stationarity of this data set was found in Sections 3.1.2, 3.2.1 and 3.3. The Dependent SiZer analysis for this time period is shown in Figure 14.

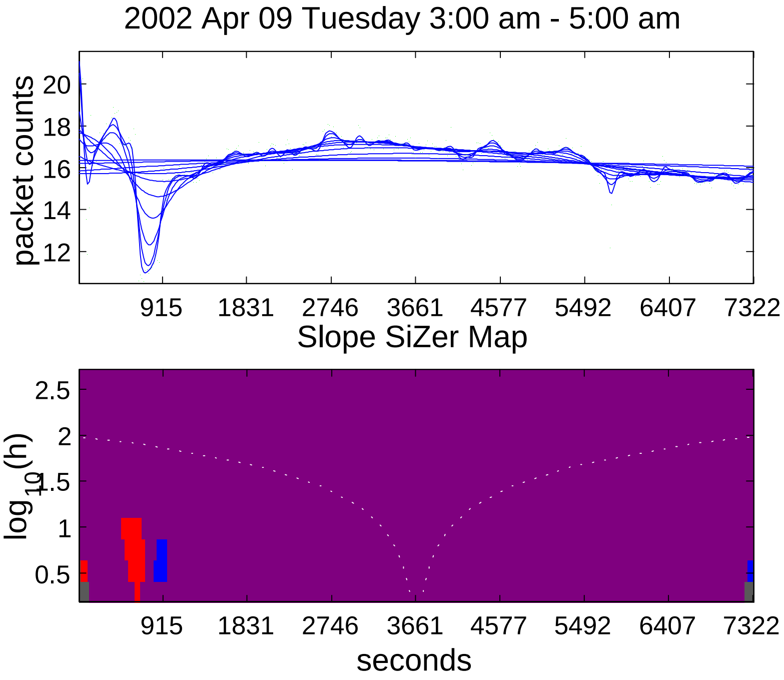

FIGURE 14: *Dependent SiZer analysis of the Tuesday Predawn UNC Link data. Shows a statistically significant valley, whose magnitude is larger than can be explained by FGN, with* $H = 0.9$*.*

The blue family of smooths in the top panel suggest a deep valley, where the overall transmission rate has dropped from around 17 packets per millisecond to about 12 packets per millisecond. The red region followed by blue, at those same times (columns), in the SiZer map shows that this valley is statistically significant, i.e. that it is inconsistent with the amount of wiggliness that is generated by the FGN model. The fact that the significant structure appears near the bottom of the SiZer map is an indication that this is a relatively small time scale phenomenon. The cause of this dropout appears to be a temporary loss of service at a nearby router which carried about one third of the total traffic through the UNC main link.

Data Set 9 - Tuesday Morning Potential non-stationarity of this data set was found in Sections 3.1.2 and 3.2.2. A conventional SiZer analysis of this time period appears in Figure 8, and shows a strong linear trend, in addition to a suggestion of an even stronger non-linear trend near the beginning. These features of the data are confirmed in the Dependent SiZer analysis shown in Park, Marron and Rondonotti (2004).

Deeper confirmation of the non-linear component of the trend, comes from applying Dependent SiZer to the detrended data from this time series, as in Figure 15.

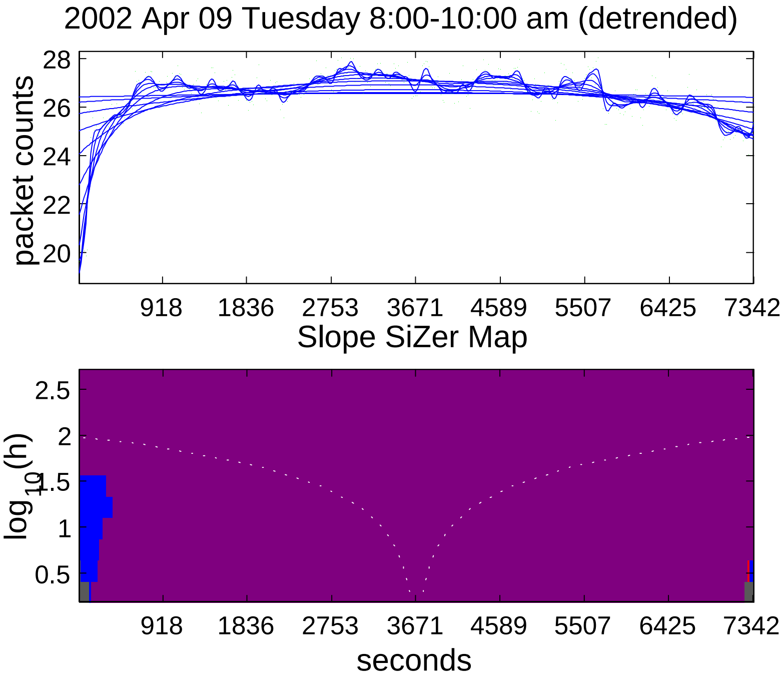

Figure 15: *Dependent SiZer analysis of the Detrended Tuesday Morning UNC Link data. Shows that the beginning increase is a statistically significant (versus FGN with $H = 0.9$) non-linear trend.*

The blue region at the left edge of this SiZer map confirms that the strong early increase in the family of smooths is statistically significant. Because this is found for the detrended data, this is clearly not part of the linear trend.

Data Set 16 - Thursday Afternoon Potential non-stationarity of this data set was found in Sections 3.1.2 and 3.3. Additional evidence was provided by Park, et al (2004a), where it was seen that the wavelet based confidence interval for H was unusually large, $[0.50, 1.09]$. The Dependent SiZer map is not shown here (but is web available) for this time period, because it does not flag significant structure. But the unusual Hurst parameter estimation process encountered above has motivated a much more careful analysis of this time period.

A first version is available as a movie file, available from the link "Thu 1300 April 11 2002" in the bottom block, "Comparison of 2002 and 2003 UNC Subtraces" on the web page Park (2004). Similar movies for some other time periods are also available there. This shows a moving window version of both Hurst parameter estimation and Dependent SiZer analysis. This suggests that the non-stationarity is a small time scale (only 8 seconds) drop out, where the traffic nearly stopped completely, suggesting that the main link itself was down briefly, or else a router that was directly attached and carried nearly all the traffic. This moving window analysis evolved into Local Analysis of Self-Similarity, as proposed in Stoev, et al (2004b), where this same time period provides a central example.

Another approach to finding the non-stationarity present for this time period, which combines SiZer analysis with a wavelet decomposition, can be found in Park, et al (2004a). In that paper (where the data were analyzed at the 10 ms scale), it was verified that the dropout caused the non-stationarity, by deleting that time interval, and connecting the remaining pieces into a single time series. The Hurst parameter estimate of the modified series was now in the same region as for most of the other data sets, $H = 0.90$, and the Confidence Interval was also more typical $[0.88, 0.92]$.

An important lesson of all of these analyses is that non-stationarities can be quite local in time, and that for most time periods, the conventional FGN models can fit very well.

Data Set 22 - Saturday Afternoon Potential non-stationarity of this data set was found in Sections 3.1.2, 3.2.1 and 3.3. Dependent SiZer analysis of this time period reveals a large, relatively long lasting (6 minutes) period of unusually heavy traffic. The analysis is not shown here, because it appears in Park, Marron and Rondonotti (2004) (and is web available at Park 2004). In the analysis of that paper, some careful zooming in on the region of interest suggests that this non-stationarity was caused by an IP port scan. Again it was verified that this unusual behavior caused the apparent non-stationarity, by deleting that time segment, and connecting the remaining pieces into a single series, which exhibited stationary behavior.

4.2 Whittle Diagnostic Plot Analysis

As noted above, the wavelet spectrum, and several enhancements, provide detailed diagnostics of the H estimation process. Similar diagnostics are available for Local Whittle estimation, as shown in Figure 16.

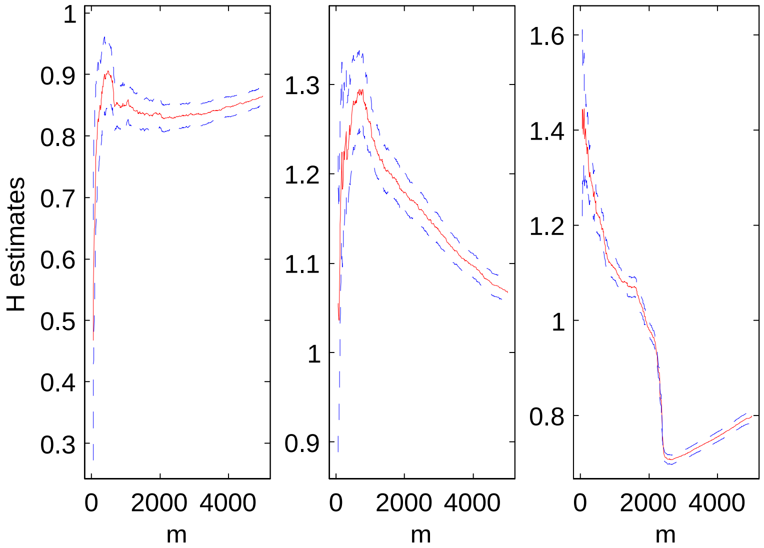

Figure 16: *Local Whittle Diagnostic plot, showing estimated Hurst parameter as a function of the window parameter m. For (a) Laboratory Data, showing very stable behavior, (b) Tuesday Predawn, showing typical real data behavior (c) Saturday afternoon, indicating Hurst parameter estimation is very unstable, consistent with the above analyses.*

Each panel of Figure 16, shows the Local Whittle Estimate, as a function of the window parameter m, which determines the number of terms in the summation of (5).

The left hand plot is based on the laboratory data with throughput 20 Mbps, and shows unstable estimates of H for small m, but they stabilize around $m = 700$ and thereby remain in the range 0.83–0.86 with a standard error in the range .01–.02. Thus, for a series such as this we can feel fairly confident in our estimate of H. All five lab data series produce plots similar to these, indicating stable estimates of H. In contrast, the real data produce a variety of plots, some of which also indicate stable H while others look more like the middle plot in Figure 16 (for Tuesday Predawn). This plot shows high variability that does not appear to stabilize as m increases. For this series, the choice of m and hence the estimate of H seems highly subjective. The third plot in Figure 16 is for Saturday Afternoon, and shows a very unusual and unstable case. In this case it is worth noting that the confidence bands are very narrow in comparison with the variation in H itself, implying that the confidence bands cannot be trusted.

For example, in the three series of Fig. 16, the best values of m were respectively 3300, 890, 4900 with corresponding point estimates and 95% confidence intervals of 0.84 (0.82–0.86), 1.27 (1.24–1.31) and 0.758 (0.753–0.763). In the first case this seems very consistent with the overall stable appearance of the plot, but in the latter two cases, the tuned method may very well be construed as leading to misleadingly precise estimates. For the Tuesday Predawn data (center panel), the large point estimate of $H > 1$ itself suggests nonstationarity, but with the Saturday Afternoon data, the main indicator that something is wrong is the shape of the plot itself.

4.3 Conclusions about non-stationarity

These examples strongly suggest that, at least in the context of Internet traffic, blind estimation of the Hurst parameter can be misleading. Careful analysis of situations where Hurst parameter estimates are larger than 1 suggests that an important cause of this is gross non-stationarities. But the nature of these is quite different from what would be expected from a classical non-stationary Gaussian process such as Fractional Brownian Motion. Instead the non-stationarities tend to appear as either long time scale linear trends, or as relatively short time scale major dropouts or spikes, of a magnitude that is far larger than expected under FGN. However, for most time periods (even inside those time periods containing clear non-FGN features) the FGN model is actually quite consistent with the data.

Thus even though a number of estimated Hurst parameters are larger than 1, which is inconsistent with mathematical assumptions underlying the whole process of Hurst parameter estimation, the exercise is still worthwhile. This is because the data are consistent with stationary FGN *most of the time*, and the time periods where there are serious departures are thus clearly flagged by the estimated H being bigger than 1. This has been additionally verified in some cases, including data sets 16 and 22, by removing clear non-stationary segments, and verifying that the remainder behaves in a stationary way. A different approach to adjusting non-stationary parts of time series, which resulted in a stationary series, was taken by Shen, Zhu and Lee (2004).

In the next section we make some suggestions for realistic simulations of Internet traffic.

4.4 Improved Simulation

The lessons learned in this analysis suggest potential improvements in both purely simulated, and in laboratory generated, experiments.

First off, since the FGN gave a reasonable fit over many time intervals, it should be the baseline for simulated data. Since the lab data was generally consistent with the FGN, current procedures are seen to be effective for most time intervals.

Second, many time periods exhibit significant linear trends, so it may make sense to build these into either type of model. This is easy to do in either case.

Third, there are occasional features in the real data that are far different than such models can generate. In particular, no simple stochastic model seems capable of containing the wide array of anomalies exhibited by the data. Unfortunately these anomalies occur with rather low probability, so it is likely that we have not seen all types, and it is hard to suggest how such features should be added into models. Ideally more experience could be gained through the collection and analysis of more data. In the meantime, existing models could be modified to include features (dropouts and huge spikes) of the type that have been encountered to date, depending on the specific goal of the experiment.

5 Acknowledgement

This work has been performed over several years by the various authors. Most progress was made during the SAMSI Program on Network Modelling for the Internet. The work of J. S. Marron was also partially supported by NSF Grants DMS-9971649 and DMS-0308331. The work of R. L. Smith was partially supported by NSF Grant DMS-0084375.

References

[1] Abry, P. and Sellan, F. (1996) The wavelet-based synthesis for fractional Brownian motion proposed by F. Sellan and Y. Meyer: remarks and fast implementation, *Applied and Computational Harmonic Analysis*, 3, 377-383.

[2] Abry, P. and Veitch, D (1998) Wavelet analysis of long range dependent traffic. *Trans. Info. Theory*, 44, 2–15.

[3] Bardet, J.-M., Lang, G., Moulines, E. and Soulier, P. (2000) Wavelet estimator of long-range dependent processes, *Statistical Inference for Stochastic Processes* 3, 85-99.

[4] Bardet, J.-M. (2002) Statistical study of the wavelet analysis of fractional Brownian motion, *IEEE Transactions on Information Theory*, 48, 991-999.

[5] Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman & Hall.

[6] Brockwell, P. J. and Davis, R. A. (1996) *Time Series: Theory and Methods*, Springer, New York.

[7] Cao, J., Cleveland, W. S., Lin, D., and Sun, D. X. (2001) On the Nonstationarity of Internet Traffic, *Proc. ACM SIGMETRICS '01*, 102-112.

[8] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807–823.

[9] Chilès, J. P. and Delfiner, P. (1999) *Geostatistics: modeling spatial uncertainty*, Wiley, NewYork.

[10] Crovella, M. E. and Bestavros, A. (1996) Self-similarity in World Wide Web traffic evidence and possible causes, *Proceedings of the ACM SIGMETRICS 96*, pages 160–169, Philadelphia, PA.

[11] Danzig, P., Mogul, J., Paxson, V. and Schwartz, M. (2000) Web Site: http://ita.ee.lbl.gov/html/contrib/fft-fgn.html.

[12] Daubechies, I. (1992) *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics.

[13] Dietrich, C. R. and Newsam, G. N. (1997) Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix, *SIAM Journal on Scientific and Statistical Computing*, 18, 1088–1107.

[14] Doukhan, P., Oppenheim, G., and Taqqu, M. S., editors (2003) *Theory and Applications of Long-Range Dependence*. Birkhäuser.

[15] Feldmann, A. Gilbert, A. C. and Willinger, W. (1998) Data networks as cascades: investigating the multifractal nature of Internet WAN traffic, *Computer Communication Review, Proceedings of the ACM/SIGCOMM '98*, 28, 42–55.

[16] Floyd, S. (2004) Pointers to the literature: congestion control, Internet available at http://www.icir.org/floyd/.

[17] Garrett, M. W. and Willinger, W. (1994) Analysis, Modeling and Generation of Self-Similar Video Traffic, *Proc. of the ACM SIGCOMM'94*, London, UK, 269-280.

[18] Geweke, J. and Porter-Hudak, S. (1983) The estimation and application of long-memory time series model, *J. Time Series Anal.*, 4, 221–238.

[19] Hannig, J. and Marron, J. S. (2004) Advanced distribution theory for SiZer, unpublished manuscript, Internet available at: http://www.stat.unc.edu/postscript/papers/marron/SiZerSimul/SiZerDist8.pdf

[20] Henry, M. and Robinson, P. M. (1996) Bandwidth choice in Gaussian semiparametric estimation of long-range dependence, in *Athens Conference on Applied Probability and Time Series Analysis, Volume II: Time Series Analysis, In memory of E.J. Hannan*, ed. by P. M. Robinson, and M. Rosenblatt. Springer Verlag, New York, 220–232.

[21] Hernández-Campos, F., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2004) Variable Heavy Tails in Internet Traffic, tentatively accepted by *Performance Evaluation*.

[22] Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer*, 1, 69–91.

[23] Kim, M. and Tewfik, A. H. (1992) Correlation structure of the discrete wavelet coeffcients of fractional Brownian motion, *IEEE Transactions on Information Theory*, 38, 904-909.

[24] Kunsch, H. R. (1987) Statistical aspects of self-similar processes, In *Proc. 1st World Congress Bernoulli Soc.* (Yu. Prohorov and V.V. Sazonov, eds.) 1, 67–74, Utrecht, VNU Science Press.

[25] Kurose, J. F. and Ross, K. W. (2004) *Computer Networking: A Top-Down Approach Featuring the Internet*, third edition, Pearson Addison Wesley.

[26] Le, L. Aikat, J. Jeffay, K. and Smith, F. D. (2003) The Effects of Active Queue Management on Web Performance, *Proc. of ACM SIGCOMM 2003*, 265-276.

[27] Le, L., Hernández-Campos, F. and Park, C. (2004) Web site: http://www-dirt.cs.unc.edu/net_lrd/.

[28] Mandelbrot, B. and Van Ness, J. (1968) Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4), 422-437.

[29] Mandelbrot, B. B. (1969) Long-run linearity, locally Gaussian processes, H-spectra and infinite variance, *International Economic Review*, 10, 82–113.

[30] Marron, J. S., Hernández-Campos, F. and Smith, F. D. (2002) Mice and Elephants Visualization of Internet Traffic, in *COMPSTAT 2002 - Proceedings in Computational Statistics - 15th Symposium held in Berlin*, eds. Härdle, W. and Rönz, B., Physika Verlag, Heidelberg. Internet available at: http://www.cs.unc.edu/Research/dirt/proj/marron/MiceElephants/.

[31] Marron, J. S., Hernández-Campos, F. and Smith, F. D. (2004) A SiZer analysis of IP flow start times, to appear in *Proceedings of Conference in Honor of Erich Lehmann*, Institute of Mathematical Statistics.

[32] Park, C. (2004) Web Site: http://www-dirt.cs.unc.edu/net_lrd/DepSiZer/SiZER_View.html.

[33] Park, C., Godtliebsen, F., Taqqu, M., Stoev, S. and Marron, J. S. (2004a) Visualization and inference based on wavelet coefficients, SiZer and SiNos, submitted to *Journal of Computational and Graphical Statistics*.

[34] Park, C., Hernández-Campos, F., Marron, J. S. and Smith, F. D. (2004b) Long-Range-Dependence in a Changing Internet Traffic Mix, under revision.

[35] Park, C., Marron, J. S. and Rondonotti, V. (2004) Dependent SiZer: goodness of fit tests for time series models, to appear in *Journal of Applied Statistics*.

[36] Paxson, V. (1994) Empirically-Derived Analytic Models of Wide-Area TCP, Connections. *IEEE/ACM Transactions on Networking*, 2, 316–336.

[37] Paxson, V. (1995) Fast Approximation of Self-Similar Network Traffic, Technical report LBL-36750/UC-405.

[38] Paxson, V. and Floyd, S. (1995) Wide Area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, 3, 226–244.

[39] Pipiras, V. (2003) Wavelet-based simulation of fractional Brownian motion revisited, Preprint.

[40] Resnick, S. and Samorodnitsky, G. (1999) Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues, *Queueing Systems*, 33, 43–71.

[41] Riedi, R. and Willinger, W. (1999) Toward an improved understanding of network traffic dynamics, in *Self-similar Network Traffic and Performance Evaluation,* Wiley, New York.

[42] Robinson, P. M. (1995) Gaussian semiparametric estimation of long range dependence, *The Annals of Statistics*, 23, 1630–1661.

[43] Roughan, M. and Veitch, D. (1999) Measuring Long-Range Dependence under Changing Traffic Conditions, *IEEE INFOCOM'99*, 1513–1521.

[44] Shen, H., Zhu, Z. and Lee, T. (2004) Robust estimation of self-similarity parameter in network traffic using wavelet transform, unpublished manuscript.

[45] Sherman, B., Willinger, W. and Teverovsky,V. (2000) Web site: http://math.bu.edu/people/murad/methods/var/.

[46] Smith, F. D. Hernandez Campos, F. Jeffay, K. and Ott, D. (2001) What TCP/IP Protocols Headers Can Tell Us About The Web, *Proc. of ACM SIGMETRICS 2001*, 245-256.

[47] Stoev, S., Taqqu, M., Park, C. and Marron, J. S. (2004a) On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic, submitted to *Computer Networks.*

[48] Stoev, S., Taqqu, M., Park, C., Michailidis, G. and Marron, J. S. (2004b) LASS: a tool for the local analysis of self-similarity in Internet traffic, submitted to *Computational Statistics and Data Analysis.*

[49] Taqqu, M. and Levy, J. (1986) Using renewal processes to generate LRD and high variability, in: *Progress in probability and statistics*, E. Eberlein and M. Taqqu eds. Birkhaeuser, Boston, 73–89.

[50] Taqqu, M. S., Teverovsky, V., and Willinger, W. (1995) *Estimators for long-range dependence: An empirical study.* Fractals, 3(4):785– 798.

[51] Taqqu, M. S., Teverovsky, V. (1998) On Estimating the Intensity of Long-Range Dependence in Finite and Infinite Variance Time Series, In R. J. Alder, R. E. Feldman and M.S. Taqqu, editor, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, 177–217. Birkhauser, Boston.

[52] Tukey, J., and Tukey, P. (1990) *Strips Displaying Empirical Distributions: Textured Dot Strips.* Bellcore Technical Memorandum.

[53] Veitch, D. (2002) Web site: http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html.

[54] Veitch, D., Abry P., and Taqqu, M. (2003) On the automatic selection of the onset of scaling. *Fractals*, 11, 377-390.

[55] Veitch, D. and Abry, P. (1999) A wavelet based joint estimator for the parameters of LRD, in *Special issue on Multiscale Statistical Signal Analysis and its Applications, IEEE Transactions on Information Theory*, 45, 878-897.

[56] Willinger, W. Taqqu, M. S. Sherman, R. and Wilson D. (1997) Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level, *IEEE/ACM Transactions on Networking*, 5:71–86.

[57] Wood, A. T. A. and Chan, G. (1994) Simulation of stationary Gaussian processes in $[0,1]^d$, *Journal of Computational and Graphical Statistics*, 3, 409–432.