Long-range dependence and heavy-tail modeling for teletraffic data

Olivier Cappé `<cappe@tsi.enst.fr>`, Eric Moulines `<moulines@tsi.enst.fr>`,
Jean-Christophe Pesquet `<pesquet@univ-mlv.fr>`,
Athina Petropulu `<athina@cbis.ece.drexel.edu>`, Xueshi Yang `<reeves@iason.ece.drexel.edu>`

January 24, 2002

1 Introduction

Analysis and modeling of computer network traffic is a daunting task considering the amount of available data. This is quite obvious when considering the spatial dimension of the problem, since the number of interacting computers, gateways and switches can easily reach several thousands, even in a Local Area Network (LAN) setting. This is also true for the time dimension: W. Willinger and V. Paxson in [42] cite the figures of 439 million packets and 89 gigabytes of data for a single week record of the activity of a university gateway in 1995. The complexity of the problem further increases when considering Wide Area Network (WAN) data [28].

In light of the above, it is clear that a notion of importance for modern network engineering is that of *invariants*, *i.e.* characteristics that are observed with some reproducibility and independently of the precise settings of the network under consideration. In this tutorial paper, we focus on two such invariants related to the time dimension of the problem, namely, long-range dependence, or self-similarity, and heavy-tail marginal distributions. Both characteristics arise in most "scalar signals" that can be extracted from complete network traffic traces [27, 35, 3, 43]. Typical "scalar signals" include continuous-time point processes constructed from recording the arrival times of successive IP (Internet Protocol) packets at some point of the network, or a time series obtained by counting the size of the data transferred during some time intervals.

In order to illustrate and motivate the technical part of this tutorial, we begin with a statistical description of the traffic data that will be considered throughout the paper. A striking feature, which corroborates the conjecture that long-range dependence and heavy-tailness are really meaningful traffic invariants, is that they can be observed, to some extent, without using any specific experimental pro-
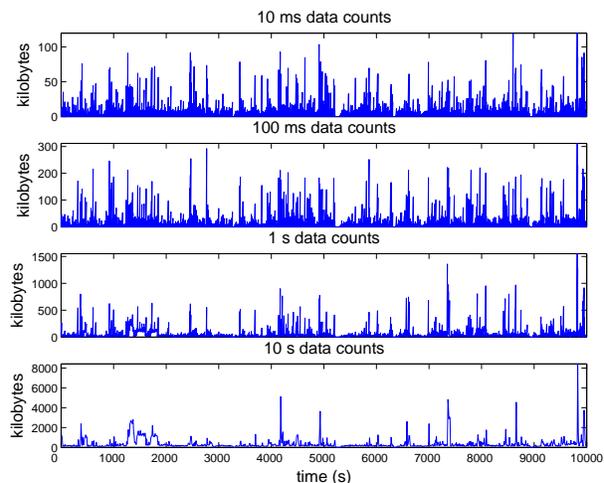

Figure 1: The Drexel data (10 000 s in total) viewed through four different aggregation intervals: from top to bottom, 10 ms, 100 ms, 1 s and 10 s.

tocol. Accordingly, the traffic data shown in figure 1 corresponds to actual 100 Mbps Ethernet traffic, which was measured on a WWW/Email/FTP/-Computing server located in the ECE Department of Drexel University. It was captured using the `Snoop` program, which is part of the Sun Solaris operating system. To generate this trace, all packets of private connections with this server, broadcasting, and multi-casting were captured and time-stamped during several hours. In the following, we only consider byte counts (size of the transferred data) measured on 10 ms intervals, which is the data represented in the top plot of figure 1. The overall length of the record is about three hours (exactly, 10^4 seconds). The three other plots in figure 1 correspond to the "aggregated" data obtained by accumulating the data counts on increasing time intervals. The striking feature in figure 1 is that the aggregation is not really successful in smoothing out the data. The aggregated traffic still appears bursty in the bottom plot despite the fact that each point in it is

1

obtained as the sum of one thousand successive values of the series displayed in the top plot of figure 1. Similar characteristics have been observed in many different experimental setups, including both LAN and WAN data (e.g. [27, 35, 3] and the references therein).

inter–frames variance normalized by frame size (logscale)
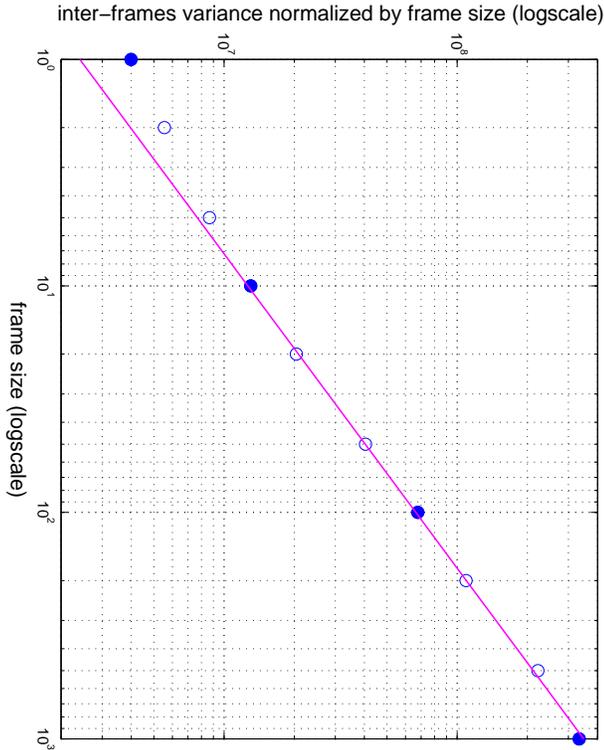
frame size (logscale)

Figure 2: Variance-time plot: Empirical variance of the aggregated process normalized by the size of the aggregation frame, plotted as a function of the size of aggregation frame expressed in 10 ms units (logarithmic scale on both axes). The dark circles correspond to the point obtained from the four plots in figure 1 while the other points correspond to intermediate aggregation intervals.

This counterintuitive behavior is clearly illustrated by figure 2, which represents the empirical variance of the aggregated data normalized by the aggregation interval as a function of the aggregation interval. Based on conventional stationary time series theory, one would expect that, for sufficiently large aggregation intervals, these points should line up along a horizontal line (or, in other words, the variance of the aggregated data should be proportional to the aggregation interval) [4]. Figure 1 clearly shows that is far from being true for the considered traffic data, since the least-squares line fitted to the rightmost six points has a slope of 0.72. Taking into account that the plot has logarithmic scales on both axes, this means that, for large ag-

gregation intervals, the variance of the aggregated process grows as the aggregation interval raised to the power 1.72. Section 2 shows that this behavior can be explained by a very slow decay of the autocovariance function, a phenomenon which we will refer to as *long range dependence*. Means for measuring and quantifying this effect, of which figure 2 is only a very basic example, are covered in detail in section 5.

probability (logscale)

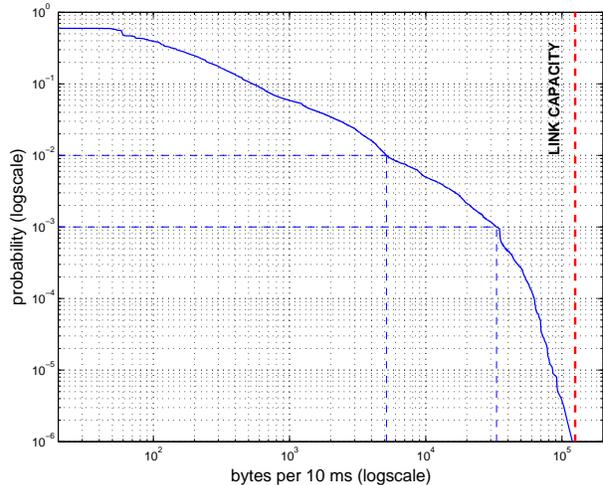LINK CAPACITY

bytes per 10 ms (logscale)

Figure 3: Empirical complementary distribution function for the Drexel data (logarithmic scale on both axes). The vertical dashed bold line at $1.25 \, 10^5$ bytes / 10ms (that is 1Mbits / 10 ms) corresponds to the capacity of the network.

The other important characteristic of the data shown in figure 1 is its extreme variability, or impulsiveness. Figure 3 illustrates this point by plotting the empirical complementary distribution function (fraction of the data larger than a given value) estimated from the data shown in the top plot of figure 1 (that is, from one million 10 ms byte counts). Starting from the leftmost part of figure 3, we observe that the plot declines from a level of 60%, which is attributable to the fact that about 40% of the data is zero. For moderately active networks like the one under consideration, there is thus a significant probability that not a single packet will be transmitted during the 10 ms interval selected as our time unit. Looking at the other end of figure 3, it is clear that the highest data sizes one actually observes corresponds to the full capacity of the network link, which is 1 Mbits / 10 ms. In between these extremes, the empirical complementary distribution function has a very slow decay. Similar characteristics have been observed in various settings [24] [22] [43], [32]. In general, impulsiveness is

dominant in data flows generated by a single user (so called source-level traffic), and LAN traffic often appears to be impulsive at both single user and multiple users levels.
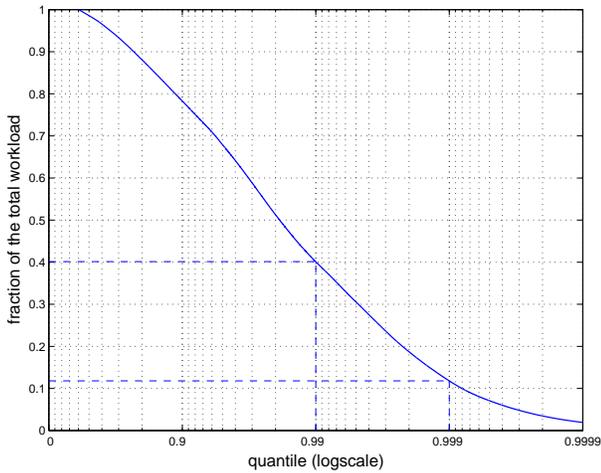

Figure 4: Fraction of the overall workload represented by data counts larger than a given quantile value (logarithmic scale on the x axis). The light dashed lines corresponds to the two quantiles (1% and 0.1%) also highlighted in figure 3. For instance, the 1% upper quantile, corresponding to transfer sizes of $5.15 \, 10^3$ bytes per 10 ms and above, accounts for about 40% of the overall workload.

Of course, one may wonder whether modeling the tail behavior, *i.e.* the appearance of very rare events, is indeed a sensible thing to do. Figure 4 shows that this is truly the case since a large fraction of the *workload*, defined as the overall size of the data transferred during the 10 000 s of traffic shown in figure 1, is due to these rare but large transfer sizes. Thus, the issue of extreme quantiles behavior and impulsiveness has an important practical impact, which justifies the need for specific models to be described in section 3.

2   The low-frequency behavior: long memory

2.1   Long range dependence

Let $\{Y_k\}_{k\in\mathbb{Z}}$ be a discrete-time second-order stationary process, with autocovariance function $\gamma(\tau) := \mathrm{cov}(Y_{k+\tau}, Y_k)^1$. Most standard time-series models assume that the covariance sequence is ab-

---

[1]In the following the index set of signals will be indicated only when there could be some ambiguity. In other words, $\{Y_k\}$ is our default notation for a discrete-time process.

solutely summable, $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| < \infty$, meaning that the dependence among the observations diminishes fast. For example, the correlation function of an ARMA process decreases geometrically fast as the correlation lag approaches infinity. Such processes are said to be Short-Range Dependent (SRD).

To model phenomena exhibiting dependence upon larger time-scale, we will consider in the sequel processes for which $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| = \infty$, and we will call them *long-memory*, or *long-range dependent* [2]. The non-summability of the autocovariance captures the intuition behind long-range dependence (LRD); even though the high-lag autocorrelations are individually small, their cumulative effect is of importance, thus giving rise to a behavior which is markedly different from that of processes with short-range dependence.

Though it is possible to develop a whole theory from this starting point, we will restrict our attention to *fractional processes*, which is the class of processes for which the autocovariance function decays according to:

$$\gamma(\tau) = L(\tau)\tau^{-2(1-H)} \quad \text{with } 1/2 < H < 1. \quad (1)$$

In (1), the function $L(\tau)$ is *slowly-varying* at infinity, *i.e.,* for all positive x, $\lim_{\tau\to\infty} L(\tau x)/L(\tau) = 1$ (typical such slowly varying functions are constant functions or ratios of two polynomials with identical degree). The coefficient H is referred to as the Hurst parameter, by reference to the hydrologist Hurst, who formally introduced these models for river flows in the 50's [19]. Equivalently, the asymptotic decay of the correlation may be characterized by the so-called "fractional index" of the process, defined as: $d \overset{\triangle}{=} H - 1/2$ (that is $0 < d < 1/2$).

The spectral density of a LRD fractional process verifying (1) is given by

$$f(\lambda) \sim |1 - e^{i\lambda}|^{-2H+1} L(1/\lambda), \quad \lambda \to 0^+ \quad (2)$$

where the function $L(\lambda)$ is regularly-varying at infinity, and the symbol $\sim$ indicates that the ratio of the left and the right hand side tends to 1. The spectral density is unbounded at zero, while near the origin obeys a power-law with index directly related to H. In contrast, the spectral density of a short-range dependent process is bounded.

2.2 Self-similarity and Long Range Dependence

Long-memory processes are intimately related to self-similar processes [30], and the two words are sometimes (improperly) used exchangeably in the network literature. A continuous time process $\{X(t)\}_{t\in\mathbb{R}}$, is self-similar with index $H > 0$, if for all $a > 0$, for any $n \geq 1$ and any n-tuple $(t_1, t_2, \cdots, t_n)$, the sequences $(X(at_1), \cdots, X(at_n))$ and $(a^H X(t_1), \cdots, a^H X(t_n))$ have identical distributions (see also companion article [1] by Abry, Baraniuk, Flandrin, Riedi and Veitch in this issue). Thus, self-similarity implies that a change of the time scale is equivalent to a change in state space scale (this definition does of course require that the process be defined in continuous time since we must be able to scale the time axis by any positive factor a). The term *self-similar* was coined by Mandelbrot and is now standard. A process $\{X(t)\}_{t\in\mathbb{R}}$ has *stationary increments* if, for all $t_0 \in \mathbb{R}$, the shifted processes $\{X(t + t_0) - X(t_0)\}_{t\in\mathbb{R}}$ have identical distributions, irrespectively of the value of t_0. A process $\{X(t)\}_{t\in\mathbb{R}}$ is said to be *H-sssi* if it is self-similar with stationary increments. A beautiful result is that there is a unique Gaussian H-sssi process, $\{B_H(t)\}_{t\in\mathbb{R}}$, the *fractional Brownian motion* or *fBm* for short. There are also *non-Gaussian finite-variance H-sssi* processes, but none of them play the same prominent role as the fBm, which has been used successfully to model a variety of natural phenomena, such as terrains, coast lines, and clouds, and, of course, network traffic data. Note that, if $H = 1/2$, the fBm coincides with the classical Brownian motion [38].

Since fractional Brownian motion has stationary increments, its sampled increments, $Y_H(k) = B_H(k) - B_H(k-1)$, $k \in \mathbb{Z}$, form a stationary sequence. The sequence $\{Y_H(k)\}_{k\in\mathbb{Z}}$ is called *fractional Gaussian noise* (FGN). Note that, as $\tau \to \infty$, the autocovariance function of the FGN is $\gamma(\tau) \sim \sigma_0^2 H(2H-1)\tau^{2H-2}$ (for $H \neq 1/2$), and thus, for $1/2 < H < 1$, the FGN is a long-memory fractional process, outlining the links between H-sssi processes and LRD processes.

A striking feature of FGN with $H \neq 1/2$ is that it provides a counterexample to the usual central limit theorem. Indeed, for $d_m^{-1} \sum_{k=1}^{m} Y_H(k)$ to converge in distribution, as $m \to \infty$, to a non-trivial limit, one cannot choose $d_m \sim \sqrt{m}$ but rather $d_m \sim m^H$: Since $\{Y_H(k)\}_{k\in\mathbb{Z}}$ is the increment process, $m^{-H} \sum_{k=1}^{m} Y_H(k) = m^{-H} B_H(m)$ which has the same distribution as $B_H(1)$ due to self-similarity of $\{B_H(t)\}_{t\in\mathbb{R}}$. Since $H > 1/2$, this means that a salient feature of long memory processes, such as the FGN, is that the variance of time averages decreases far less rapidly than observed in usual (short memory) models.

2.3 Aggregation, long-memory and asymptotic self-similarity

For discrete time processes, self-similarity is better described by means of distributional invariance upon aggregation and scaling. More precisely, let $\{Y_k^{(m)}\}_{k\in\mathbb{Z}}$ denote the aggregated process of order m :

$$Y_k^{(m)} := \sum_{j=(k-1)m+1}^{km} Y_j, \quad k = 1, 2, \ldots$$

that is, the process defined as the sum over consecutive blocks of size m of the original process $\{Y_k\}$. A process is *exactly self-similar with index* α if, for all $m \geq 1$, the scaled aggregated process $\{m^{-\alpha} Y_k^{(m)}\}_{k\in\mathbb{Z}}$ has the same distribution as $\{Y_k\}$. It is easily seen that FGN with Hurst parameter H is self-similar with index H, and it can be shown that it is the only discrete-time finite-variance process that has this property.

A weaker form of discrete-time self-similarity can be defined by examining the limiting behavior of $\{Y_k^{(m)}\}_{k\in\mathbb{Z}}$ for large values of m. Recall that if $\{Y_k\}$ was a sequence of zero-mean i.i.d random variable with finite variance σ^2, the central limit theorem would imply that, at the limit of large aggregation m, the finite dimensional distributions of $\{m^{-1/2} Y_k^{(m)}\}_{k\in\mathbb{Z}}$ converge to those of a Gaussian white noise with variance σ^2. This is of course no longer the case for long memory processes: Assume that $\{Y_k\}$ is a stationary fractional process with index $1/2 < H < 1$, a simple but profound result of the theory is that under mild conditions, as $m \to \infty$, the limiting distribution of $\{m^{-H} Y_k^{(m)}\}_{k\in\mathbb{Z}}$ is that of the FGN of index H. This fact draws strong links between LRD and self-similarity. This is the reason why LRD processes are sometimes referred to as *asymptotically self-similar*. Intuitively, the most striking feature of asymptotically self-similar processes is that their aggregated processes do not tend to second-order white Gaussian noises.

Aggregation can thus serve as a means to estimate the Hurst coefficient. The procedure, often referred under the catchy name of *variance-time plot* consists of regressing the variance v_m of $\{Y_k^{(m)}\}_{k \in \mathbb{Z}}$ on the aggregation size m. Under weak dependence conditions, v_m is linear w.r.t m at the limit of "large" aggregation intervals. If the process is LRD, the variance of the running aggregate process variance v_m increases more quickly than m. More precisely, if the process is LRD fractional with Hurst parameter H, then $v_m \propto m^{2H}$. The Hurst coefficient can thus be deduced by estimating the slope in the regression of $\log v_m$ on $\log m$. Figure 2 displays such a variance-time plot using logarithmic scales. The procedure has been often used in the early days of network traffic analysis [27], [41], though it is now well established that this is a poor estimator of the memory coefficient from a statistical perspective. These estimators are outperformed by either spectral domain (see section 5.1) or wavelet-domain estimators (described in [1]).

One of the salient features of self-similar processes is that they capture an important empirical law, commonly referred to as the *Hurst law*. Consider a finite stretch of observations, $Y_1, \ldots, Y_n$. Denote by $\bar{Y}_n$ the sample mean and S_n^2 the sample variance, and define $W_k := \sum_{\ell=1}^k Y_\ell - k\bar{Y}_n$, $k = 1, \ldots, n$. The *rescaled adjusted range statistic* (R/S statistic for short) is given by

$$Q_n := S_n^{-1}\big(\max(0, W_1, W_2, \ldots, W_n) - \min(0, W_1, W_2, \ldots, W_n)\big).$$

When Y is asymptotically self-similar with self-similarity parameter H, $\mathbb{E}[Q_n] = n^H L(n)$, where $L(n)$ is slowly varying at infinity. This property is the so called *Hurst law* which states that the range of the aggregated process is of much larger order when long term dependences are present (the appropriate asymptotic scaling factor would be $n^{1/2}$ only for short-memory processes). The R/S statistics thus provides an alternative estimator of the Hurst parameter. Hurst has found experimentally that many *naturally occurring* time series appear to satisfy the Hurst law [19]. This estimator has however rather poor statistical performance, and is nowadays seldomly used in practice.

3   Modeling impulsiveness

In the previous section, we saw that the presence of long-range correlations in the data could partly account for the unusual range of the aggregated process that was observed in figure 1. But figure 3 indicates that the marginal distribution of the data is already very dispersed, with non negligible probability of observing extremely large values. This feature can be captured through the use of so-called *heavy-tail* models which exhibit tails that decay much slower than Gaussian or exponential distributions. A perhaps surprising characteristic of some of these models is that they have infinite variance. We should emphasize here that, like most physical phenomena, traffic time series can hardly be considered as having infinite variance. Employing heavy-tail models in modeling such data only constitutes an approximation of the real tail behavior.

3.1   Heavy tail distributions

A random variable X is *regularly-varying* index α, if there exists a slowly-varying function at infinity, $L(x)$, (see the definition given in Section 2.1) such that, as $x \to \infty$,

$$\mathbb{P}(|X| \geq x) \sim \frac{L(x)}{x^\alpha}. \tag{3}$$

The variable X is said to have *heavy tails with infinite variance* if X is regularly varying with index $0 < \alpha < 2$. In those cases, the variance of X is infinite (if $\alpha < 1$ the mean itself is infinite). Mandelbrot [30] refers to (3) to as the *Noah effect* or the *infinite variance syndrome*. Intuitively, an infinite variance means that the random variable X can fluctuate far away from its central range (the reader may think of it as the median level, since this is always correctly defined whatever the value of α is). For example, the Pareto distribution with complementary distribution function:

$$P(X \geq x) = \begin{cases} (\frac{x_0}{x})^\alpha, & x \geq x_0, \\ 1, & x < x_0, \end{cases} \tag{4}$$

where x_0 is positive constant and $0 < \alpha < 2$, satisfies the *heavy-tail* condition.

3.2 α-stable distributions

A particular class of heavy-tail distributions with infinite variance are the (symmetric) α-stable ones.

Their characteristic function has the following form:

$$\Phi(\theta) = \mathbb{E}[\exp(i\theta X)] = \exp(-|\theta/\sigma|^\alpha), \quad (5)$$

where σ is a scaling parameters. Further generalizations of (5), omitted here for brevity, include introducing a shift and a skewness (asymmetry) parameters. Such a distribution is "stable", in the sense that if $X_1, \ldots X_n$ are i.i.d. random variables with distribution given by (5), then $m^{-1/\alpha} \sum_{i=1}^m X_i$ also has the same distribution. Thanks to this property, α-stable models have a number of desirable features in terms of estimation and processing [34].

4 Stochastic modeling of long-range dependence and heavy tails

After having reviewed the main concepts pertaining to long-range dependence and heavy-tail distributions, we now consider the models that have been proposed for capturing long-range dependence and heavy-tailness in the context of network traffic. One approach relies on *behavioral* models, which attempt to mimic the trends observed in measured data, without taking into account the nature of network traffic. Such models are also referred to as "black-box". The second approach relies on *structural* modeling. It attempts to explain the observed data characteristics by using knowledge about the traffic, such as the fact that it results from superposition of a large number of sources that share common resources (the individual user sessions). Behavioral modeling is often severely criticized in the network engineering literature, see [42] for instance, because the parameters of the fitted models are not related to network parameters, and thus do not lend themselves to an easy interpretation. Nevertheless, behavioral models can be useful for simulation purposes or traffic forecasting.

It is clear that one can extract a large amount of information from a traffic trace, which can be used in designing structural models. For instance, the header of the TCP packets may be used to, at least partly, recover individual sessions, end-to-end protocols, or application level information. However, it is often not possible to describe the full complexity of the data using a structural model, which means that, at some stage, one might have to resort to behavioral modeling.

4.1 Behavioral models

In this section, we only describe the linear models, which build on the familiar ARMA models, and constitute a simple but yet rich class of models for series presenting long-range dependence effects and/or heavy marginal tail properties.

Linear second-order stationary models are defined by

$$Y_k = \sum_{\ell=0}^{\infty} a_\ell Z_{k-\ell} \quad (6)$$

where $\{Z_k\}$ is a sequence of independent or identically distributed (i.i.d) zero-mean finite-variance random variables and $\sum_{\ell=0}^{\infty} a_\ell^2 < \infty$. The fact that $Y = \{Y_k\}$ is LRD implies that $\sum_{\ell=0}^{\infty} |a_\ell| = \infty$, otherwise such process has absolutely summable autocovariance sequence. $\{Y_k\}$ is a fractional process if $a_\ell \sim C\ell^{d-1}$, with $C \in \mathbb{R}$, as $\ell \to \infty$. The basic building block in defining such a model is the fractional difference operator [18], which is a fractional generalization of the classical difference operator. For $d < 1/2$, consider the formal power series expansion of $(1-z)^{-d}$,

$$(1-z)^{-d} = \sum_{j=0}^{\infty} b_j(d) z^j = \sum_{j=0}^{\infty} \frac{\Gamma(d+j)}{\Gamma(d)\Gamma(j+1)} z^j \quad (7)$$

where Γ denotes the Gamma function. The process $\{X_k\}$ is the fractional integration of order d of the process $\{Y_k\}$, if

$$X_k = (1-B)^{-d} Y_k = \sum_{j=0}^{\infty} b_j(d) Y_{k-j} \quad (8)$$

where B denotes the backward shift operator $(BY_k = Y_{k-1})$. For $d = 1, 2, \ldots$, $(1-B)^{-d}$ is merely the inverse of the difference operator $(1-B)$ iterated $|d|$ times. For a non-integer value of d, $(1-B)^{-d}$ has to be interpreted by its series expansion given by (7).

According to the terminology introduced by Granger and Joyeux (1980), $\{X_k\}$ corresponds to a FARIMA(p, d, q) model (namely, Auto-Regressive Fractionally Integrated Moving Average) when $\{Y_k\}$ is a causal invertible ARMA(p, q) process. The process $\{X_k\}$ is then asymptotically self-similar with index $H = d + 1/2$. FARIMA processes are more flexible than the simple FGN introduced in section 2.2, since they allow for simultaneous modeling of short-range and long-range behaviors (that is, respectively, of small-lag and large-lag regions of

the autocovariance function). The FGN process is specified by two parameters only (fractional index d and variance), which are not sufficient to capture a wide range of low-lag autocorrelation structures encountered in practice.

Defining infinite-variance linear models is a bit more intricate since we cannot anymore rely on the usual definition of the autocovariance function. It can, however, be shown that if $\{Z_k\}$ is an i.i.d. sequence of regularly-varying random variables with index $\alpha \in (0, 2)$, the series in (6) converges with probability 1 (that is the process $\{Y_k\}$ is well defined) provided that the coefficient of the impulse response of the filter decay sufficiently fast in the sense that $\sum_{j=0}^{\infty} |a_j|^{\delta} < \infty$, for some $\delta \in (0, \alpha)$.

A surprising fact, is that the sample autocorrelation function

$$\hat{\rho}_n(\tau) := \frac{\sum_{k=1}^{n-\tau} Y_k X_{k+\tau}}{\sum_{k=1}^{n} k_t^2}$$

provides a useful measure of dependence in this context despite the fact that $\{Y_k\}$ has non finite variance in this context. Indeed, it can be shown [6] that $\hat{\rho}_n(\tau)$ converges (in probability) for large values of n to

$$\frac{\sum_{j=0}^{\infty} a_j a_{j+\tau}}{\sum_{j=0}^{\infty} a_j^2},$$

which is nothing but the usual expression of the correlation function of a finite variance version of the model of (6). Readers interested in this topic and its applications for traffic data should refer to [37].

Alpha-stable linear models constitute an especially interesting subclass of linear models with infinite variance. Since any linear combination of a finite number of i.i.d. α-stable random variables is also an α-stable random variable (see section 3), there exists a simple characterization of linear models based on these distributions. A flexible model is provided by FARIMA with stable innovations. Basically the model is defined as in the second-order case (cf. (8)), except that the innovation sequence $\{Y_k\}$ is now an i.i.d. α-stable sequence. It can be shown that the existence of the process is then guaranteed provided that $d < 1 - 1/\alpha$ [38]. LRD arises when $d > 0$, which is possible only if $\alpha > 1$.

FARIMA models have been used with some success for traffic modeling, notably for variable-bit-rate (VBR) video traffic following [7] – see also [16]. Simulation and prediction of FARIMA processes is a non trivial task and nothing exact can be expected outside Gaussian FARIMA models, i.e., when $\{Y_k\}$

is and ARMA model with Gaussian innovation sequence. Nevertheless, approximate simulation techniques are available for infinite variance FARIMA models [25].

Taking into account that most signals encountered in network traffic engineering are inherently positive, linear models, and in particular Gaussian ones, can only be applied after pre-transforming the data. When dealing with packet data, it is common to consider the logarithm of the measurements [27]. This choice of transformation is rather ad-hoc, and effectively assumes that the marginal distribution of the underlying data is log-normal.

4.2 Structural models

In structural modeling, one tries to reproduce the observed features of the measurements, using models whose parameters are related to the traffic generating mechanism and the behavior of the main components of the network (e.g., protocols, network topology, routing strategy). This approach is by far the most popular in the network community, because contrary to the behavioral models, it provides insight into the impact of network design parameters and control strategies.

At some level of abstraction, a network can be modeled as a controlled non-linear system. Structural models can be either *open-loop* or *closed-loop*, depending on whether the network feedback on the input flows is taken into account or not. Examples of network feedback include the TCP adaptive window size mechanism (triggered by the observed round trip time delays and losses in the link), or the congestion avoidance strategy in routers. Closed loop models make more sense, but the network complexity is enormous and only rough approximations of these feedback mechanisms can be used for modeling purposes.

To give a feeling of what a structural model can be, we focus on what is perhaps the most intuitive and simple model for LAN or WAN traffic, namely, the ON/OFF model. This is an open-loop model, and was proposed in the seminal paper by Willinger *et al* [41] (see also [14]).

Based on a construction originally suggested by Mandelbrot [30], the ON/OFF model views the overall traffic as a superposition of many i.i.d. alternating ON/OFF processes (interested readers should check [13] for a related approach involving a varying number of sources).

An ON/OFF process represents traffic between a single source/destination pair. It alternates between two states: the ON state, during which the source transmits data at a constant rate; and the OFF state, during which the source is silent.

Figure 4.2 displays a schematic view of a section of the ON/OFF process. $X_1, X_2, \ldots$ are i.i.d non-negative random variables representing the duration of the ON states, and $Y_1, Y_2, \ldots$ are i.i.d non-negative random variables representing the duration of OFF states. Furthermore, the sequences $\{X_k\}_{k\geq 1}$ and $\{Y_k\}_{k\geq 1}$ are independent, regularly-varying with indices $\alpha_{\mathrm{ON}} > 1$ and $\alpha_{\mathrm{OFF}} > 1$, respectively. Hence, both distributions have finite mean, but their variance can be infinite, depending on whether the corresponding tail index is less than 2.

The On/Off model came to explain, for the first time, the self-similar behavior of traffic, which was experimentally observed by many researchers. It is basically the heavy-tail On and Off durations that lead to such behavior. This assumption is consistent with network measurements. Indeed, the OFF duration can have high variability, since some source model phenomena that are triggered by humans (e.g., HTTP sessions) have extremely long period of latency. There is also significant amount of empirical evidence suggesting that the ON duration is heavy-tail. A well established observation for instance is that the sizes of the files available on a server are heavy-tail, which implies that the transfer times for these files also have the same type of characteristics [5].

A mathematical expression for the process of figure 4.2 is the following:

$$W(t) = \sum_{n=0}^{\infty} 1_{[S_n, S_n + X_n)}(t) \quad \text{for } t \geq 0,$$

where S_k denotes the time of occurrence of the k-th ON period, and $1_{[s_1, s_2]}(t)$ is the indicator function, being defined as non-zero and equal to one only for $t \in [s_1, s_2]$. The distribution of S_0 is adjusted so that $W(t)$ is strict sense stationary (see [39] for details).

The autocovariance function of $\{W(t)\}$ may be expressed as [29]

$$\gamma_W(\tau) = \tau^{-(\min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}})-1)}L(\tau)$$

where L is slowly varying at infinity. If $\min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}})$ is less than 2, the process is long-
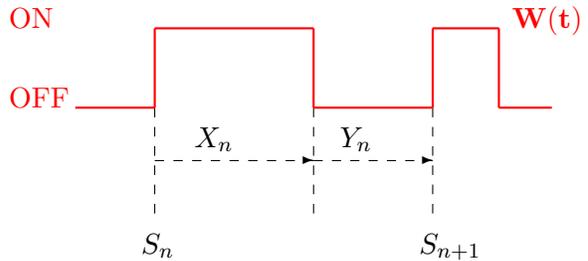

Figure 5: A single ON/OFF source

memory, with fractional differentiation index $d = 1 - \min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}})/2$.

Now consider a superposition of M i.i.d. ON/OFF sources $\{W^{(m)}(t)\}$, $m = 1, \ldots, M$. The workload process $\{N_M(t)\}_{t \in \mathbb{R}^+}$ is simply defined as the number of sources that are in the ON state at time t. $\{N_M(t)\}$ thus varies between 0 and M and also exhibits LRD when $\min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}})$ is less than 2.

Let us consider the *cumulative* input to the server between times 0 and t, defined as:

$$A_M(t) = \int_0^t N_M(s)ds.$$

It can be shown [39] that $A_M(t)$ corresponding to an increasing number of i.i.d. ON/OFF sources, under proper normalization, converges to the fractional Brownian motion, in the sense of convergence of the finite dimensional distributions. The result is formulated as a double limit, and the order of taking the limit matters (first let $M \to \infty$ and then, let $t \to \infty$). This results was thought to be of fundamental importance because it established that properly aggregated and rescaled source traffic is not only long-range dependent but also asymptotically self-similar. Recent contributions however point out that other limiting processes, which are not necessarily self-similar, can be obtained by using a different aggregation scheme (for instance, letting $t \to \infty$ and then $M \to \infty$) [31].

For practical uses of this model however, the most important fact illustrated on figure 6 is perhaps that there is a very pronounced difference of behavior depending on whether or not $\min(\alpha_{\mathrm{ON}}, \alpha_{\mathrm{OFF}})$ is smaller than 2: In the first case (red curve), the LRD phenomenon is patent with occasional very large excursions as well random very long-term trends (the x axis corresponds to 10000 points); In the second case (blue curve), one obtains a process whose oscillations are mostly high-frequency which
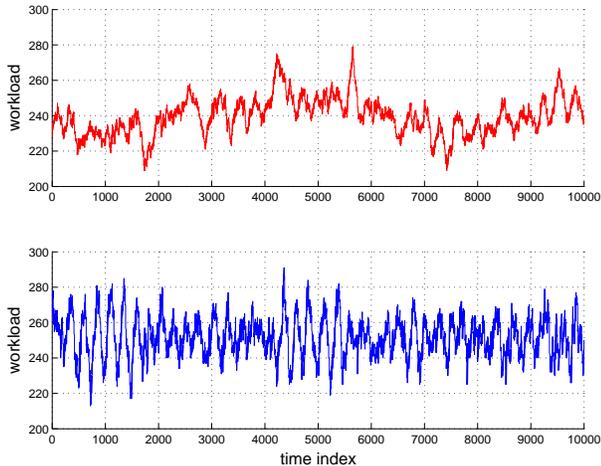
8

Figure 6: Workload resulting from the superposition of 500 sources: Top (red curve), $\alpha_{\text{ON}} = \alpha_{\text{OFF}} = 1.2$ which corresponds to a LRD limit with $d = 0.4$; Bottom (blue curve), $\alpha_{\text{ON}} = \alpha_{\text{OFF}} = 4$ which corresponds to a SRD limit.

could be represented, for instance, using a standard (short memory) autoregressive model.

Several modifications of the simple ON/OFF paradigm have been proposed. An obvious limitation of the basic ON/OFF source model is that all sources systematically emit at a nominal level. A natural way of raising this limitation consists in making the level of activity of each source random using a sequence $\{G_n\}_{n \geq 0}$ of non negative i.i.id. random variables (also called the "reward" variables). The process that describe the activity of a single source is now defined as

$$W(t) = \sum_{n=0}^{\infty} G_n 1_{[S_n, S_n + X_n)}(t) \quad \text{for } t \geq 0,$$

In [43], for instance, the author consider the case where the rewards themselves follow a Pareto distribution, with possibly infinite variance giving rise to a process which exhibit heavy-tails at the source level.

5 Quantification of long-range dependence

The application of the concepts and models that have been introduced so far for network traffic data, although dominant in the network engineering literature of the past ten years, has been somewhat controversial (compare, for example, [7] and [16], [37] and [42] or [39] and [31]). One should not be mistaken to believing that these controversies indicate a failure of such statistical approaches; they rather reflect the fact that we are here tackling a very tough problem. Estimation of the LRD index, d, or the tail index, α, from network traffic data is a statistically challenging task (let alone the issue of "verifying" experimentally whether the data is indeed self-similar). In the sequel, we illustrate some of the difficulties encountered when estimating long-range dependence parameters (this section), or tail properties (section 6) and provide an idea of the statistical performances that can be reached.

In the case of finite variance processes, LRD is described by the Hurst parameter, for the estimation of which extensive research results exist. Far fewer results exist for the quantification of LRD in the case of infinite variance processes. In this section, we provide a summary of existing literature for the above two cases.

5.1 Estimation of the Hurst parameter

Long-range dependence manifests itself either in time-domain (power-law decay of the autocorrelation) or in spectral domain (power-law singularity of the spectral density at zero frequency). Consequently, the problem of testing and estimating the long-memory behavior can be approached from a number of different angles utilizing both time-domain, frequency-domain and wavelet-domain approaches. We focus here on the frequency domain approach, and refer to [1] (this issue) for an alternative technique, based on wavelet (instead of Fourier) decomposition.

The estimation of the Hurst parameter in spectral domain amounts to estimating the exponent of the spectral density as the frequency approaches zero.

The oldest and most natural tool for spectral estimation is the tapered periodogram defined as $I(\lambda) = |d(\lambda)|^2$ where :

$$d(\lambda) = \frac{1}{\sqrt{2\pi \sum_{t=1}^{n} |h_t|^2}} \sum_{t=1}^{n} h_t X_t e^{it\lambda}$$

The use of tapers is important for the analysis of long-memory time-series in order to minimize leakage effects. For practical and theoretical reasons (see [21], [20]), it is recommended to use tapers whose Fourier transform is a finite combination of Dirichlet kernels. We restrict our attention to the tapers proposed by Hurvich and Chen [21], i.e.,

9

$h_t = (1 - \exp(i2\pi t/n))^p$, where p is referred to as the order of the kernel.

5.1.1 Local methods

The so-called local methods aim at constructing estimators that are consistent, without imposing any restrictions on the behavior of the spectral density away from zero, except from integrability on $[-\pi, +\pi]$. Recall that, since the spectral density $f(\lambda) \sim C\lambda^{-2d}$ as $\lambda \to 0^+$, then $\log f(\lambda) \simeq \log C - 2d \log(\lambda)$, *i.e.* the log spectral density is approximately linear in the log-frequency scale, and the intercept of the log-frequency is equal to (two times) the memory parameter.
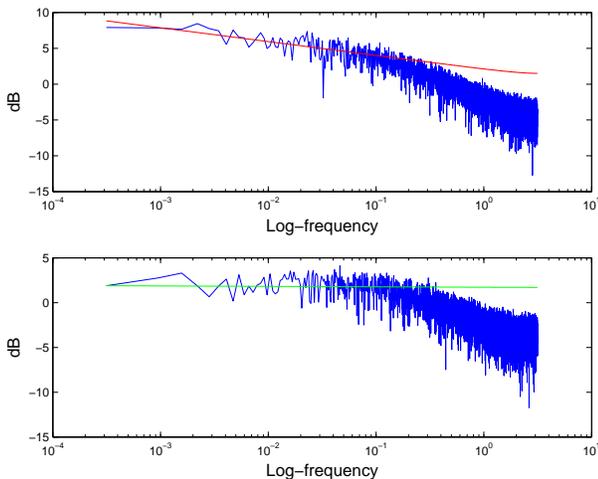

Figure 7: GPH Plot. The log-periodogram as a function of the log-frequency together with the local regression line. Top plot: Gaussian FARIMA(1,d,0), with $d = 0.4$ and $\theta = 0.9$. Bottom plot: Gaussian AR(1) process, $X_t = 0.9X_{t-1} + \epsilon_t$. In both case the number of samples is $n = 10^5$. Prior to taking the log, the periodogram ordinates has been averaged (pooled) over $m = 5$ bins.

This idea has been pushed forward in an early work by Geweke and Porter-Hudak (hence the acronym GPH) [8]. Of course, the relation $\log f(\lambda) \simeq \log(C) - 2d \log(\lambda)$ is valid only in a neighborhood of the zero frequency, and thus the regression line should be computed using only a subset of the log-periodogram ordinates $\log I(\lambda_k)$, $k \in \{1, \ldots, M\}$, where λ_k are the Fourier frequencies. The choice of the *bandwidth parameter M* is difficult problem, which critically influences the estimation procedure. The problem shares similarity with the choice of the bandwidth of a kernel for nonparametric spectral density estimation. Too small bandwidth M yields estimator with small bias (be-

cause the approximation of $\log f(\lambda)$ by $C - 2d \log(\lambda)$ is adequate for λ close to zero) but large variance (because the number of terms entering in the regression is small). A careful choice of the bandwidth should ideally mitigate these two effects. Graphical techniques [40] are very useful for selecting the bandwidth parameter based on some simple visual statistical diagnostics. Of course, this approach is only suited for off-line estimation, and automatic bandwidth selection rules are also of great practical interest. We focus here on the graphical approach for illustration purposes. A good approach is to plot $\hat{d}(M)$ as a function of M, together with the lower and upper confidence bounds, which are determined from the variance of the regression noise and the regression coefficients. To a first order approximation, this variance is independent from the memory parameter, d, and thus the bounds can be evaluated a priori (see Figure 8). Since these estimators are consistent estimators of d, it is expected that these plots have a stable regime around the true value of d.
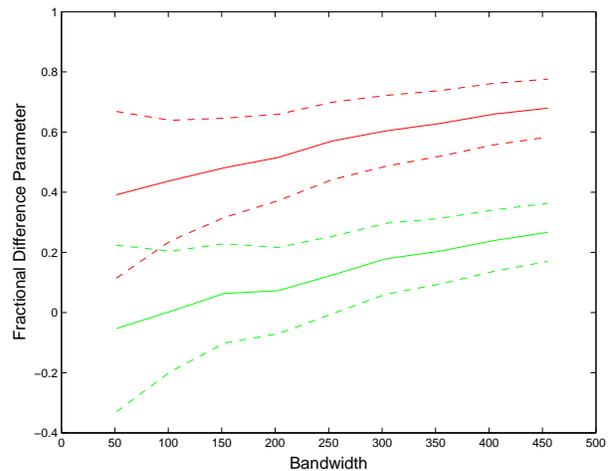

Figure 8: GPH Dynamic Plot (with 95% confidence intervals). The processes under considerations are the same as in figure 7. For small values of the bandwidth M the regression variance is very large and yields to useless estimates; for large value of M the bias dominates the variance. The red (FARIMA) and green (AR) plots appear very similar except for a shift by $d = 0.4$, showing that the major source of bias is indeed due to the approximation of the "short-memory" component.

Figure 9 shows the GPH dynamic plot corresponding to a segment of LAN traffic stretching over one-hour during the busy period. A visual inspection shows values of the memory parameter d in the range $0.35, 0.4]$, showing evidence of LRD in the data set. This corroborates the result observed
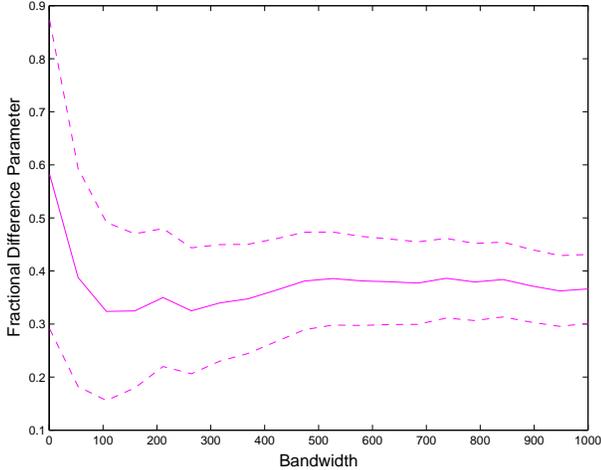
Figure 9: GPH Dynamic plot for one hour segment of the Drexel LAN traffic in a busy period (with 95% confidence intervals).

for the simple variance-time plot shown in figure 2, since we have seen in section 2.3 that the rate of growth of the variance of the aggregated process is equal to $2H = 1 + 2d$ (an thus, the 0.72 slope in figure 2, corresponds to a value of 0.36 for d).
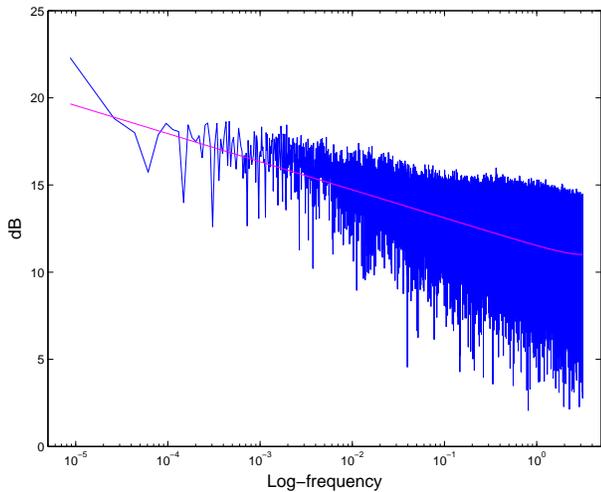

Figure 10: GPH Static plot, for a one hour segment of the Drexel LAN traffic in a busy period.

Automatic determination of the bandwidth is a difficult problem. Methods based on plug-in (requiring a pilot estimate of the spectral density of the short-memory component in a neighborhood of zero) has been discussed e.g. in [15]. Adaptive technique, based on the so-called *intersection of confidence intervals* method is presented in [9], [10] and [23].

5.1.2 Global methods

The local GPH estimator has a global counterpart. Instead of estimating d and $f^*(0)$ on a vanishing neighborhood of zero frequency, global estimators jointly estimate d and f^* over $[-\pi, \pi]$. Examples of global methods include the FEXP estimator [33] and the FAR estimator [26].

Like the GPH estimator, the FEXP estimator is based on log-periodogram regression. The principle of the FEXP estimator is to estimate simultaneously d and the coefficients of a truncated expansion of $\log f^*(\lambda) \simeq \sum_{j=0}^{q} \theta_j \cos(j\lambda)$ on the cosine basis where q is a truncation number, which plays a role similar to the bandwidth parameter, M, for local estimators. For small values of q, the bias is large (because the parametric model does not have enough degree of freedom to model the short-memory component of the spectral density) and the variance of the estimator is small (because the number of parameters to estimate is limited). On the contrary, large values of q yield to estimators with small biases but large variance. Here again, graphical methods may prove useful for off-line application. Dynamic display allows to monitor the variation of $\hat{d}$ as a function of q, helping to identify stable regions (see Figure 11). Assessment of the goodness-of-fit can then be done by choosing and fixing q and by plotting the periodogram together with the fitted spectral density. Automatic data driven selection procedure for the truncation parameter q has been considered in [23].

The FEXP estimator has been shown in [23] (see [33]) to have many interesting statistical properties. It is in general better behaved than the GPH estimator in cases where f^* is smooth because it achieves a better separation between the long-memory and the short-memory components. Comparing figures 11 and 8, the FEXP dynamic plot is less ambiguous than the GPH plot in the sense that it is easier to spot the region which correspond to a good compromise between the bias and the variance – typically for q between 10 and 15 on figure 11.

5.2 Quantification of LRD for infinite variance processes

Extensions of the previous frequency methods to LRD linear processes with infinite variance have been considered in a number of works [40]. These extensions rely on the use of the normalized peri-

odogram

$$I(\lambda) = \frac{|\sum_{t=1}^{n} h_t X_t \exp(-it\lambda)|^2}{\sum_{t=1}^{n} h_t^2 X_t^2},$$

In the absence of tapper ($h_t = 1$), $I(\lambda)$ is the Fourier transform of $\hat{\rho}_n$ the sample autocorrelation function of the considered process (remember that in section 4.1, it was shown that the sample autocorrelation function indeed converges to a finite limit for a large class of infinite variance models).

A alternative estimator of the Hurst parameter was used in [43]. It results from a linear regression applied to an empirical estimate of the logarithm of the generalized codifference [38],[36]. Despite recent advances, these estimators are still far less well-known.

6 Estimation of the tail index

The estimation of the tail index is of key importance when assessing the sporadicity of the traffic. Let $\{X_k\}_{k\geq 0}$ be a strict sense stationary process, taking on positive values only, and also let is be regularly-varying with index of variation α. The problem that we address next is the estimation of the tail index, α, from a finite stretch of data $(X_1, X_2, \ldots, X_n)$.

6.1 Hill Plot

Let $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ be the order statistics of the sample $X_1, \ldots, X_n$, which consist of the samples of the process, placed in increasing order. For some $k < n$, we define the Hill estimator [17] to be the difference between the logarithm of the k-th largest observation and the average of the logarithm of the k largest observations, i.e.,

$$\hat{\alpha}_{k,n}^{-1} = k^{-1} \sum_{j=1}^{k} \log(X_{n-k+j,n}/X_{n-k,n}).$$

When $\{X_j\}_{1\leq j\leq n}$ is an i.i.d. sample of a Pareto distribution (see eq. (4)), then the Hill estimator for $k = n$ is the maximum likelihood estimator for α^{-1}. The Hill estimator is easy to obtain, and its asymptotic behavior is well understood (at least under conditions implying short-range dependence). However, it has certain drawbacks : it is sensitive to the value of k, for the selection of which very little guidance can be offered.

The choice of k is of course a difficult issue. A useful graphical tool is to plot $\hat{\alpha}_{k,n}$ as a function
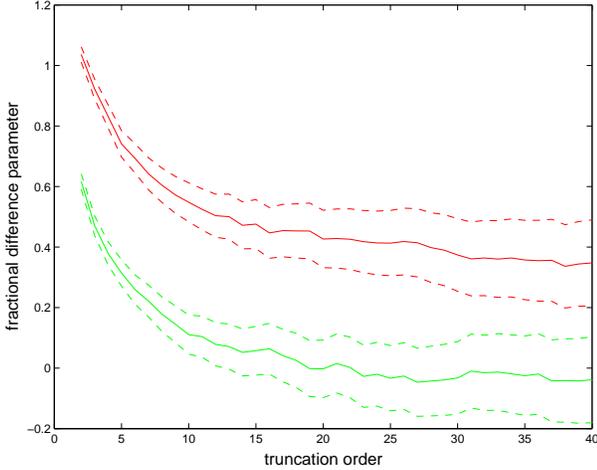

Figure 11: FEXP Dynamic Plot (with 95% confidence intervals). The processes under considerations are the same as in figure 7. For small values of the order q, the regression variance is small but the bias is very large which yields to useless estimates; for large value of q, the variance is large but the bias is small. The red (FARIMA) and green (AR) plots are very similar except for a shift by $d = 0.4$, showing that the major source of bias is indeed due to the approximation of the "short-memory" component.
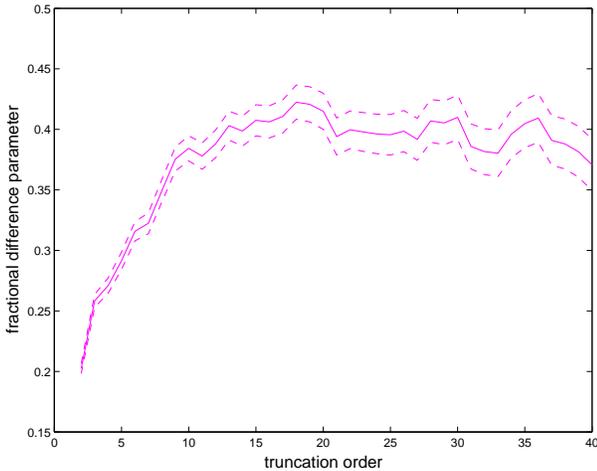

Figure 12: FEXP Dynamic plot, for a one hour segment in a busy period (with 95% confidence intervals).

of k, which yields to a *dynamic Hill plot*. Since for appropriately chosen sequence of the tail thresholds k_n, $\hat{\alpha}_{k_n,n}^{-1}$ is a consistent sequence of estimators of α^{-1}, it is expected that the Hill plot should have a stable regime at height α.

In practice, the Hill Plot exhibits extreme volatility, which makes finding a stable regime in the plot more guesswork than science. As an alternative to the dynamic Hill plot, it is sometimes useful to display the information provided by the Hill estimator as the curve $\hat{\alpha}_{[n^\theta],n}$ for values of θ between 0 and 1, where $[y]$ is the smallest integer greater or equal to $y > 0$. This alternative display is sometimes revealing, since the small order statistics get shown more clearly and cover a bigger portion of the displayed space. However, when the data is Pareto or nearly Pareto, this alternate plotting device is less useful since in the Pareto case, the Hill estimator applied to the full data set is the maximum likelihood estimator and hence the correct answer is usually found at the right end of the Hill plot.
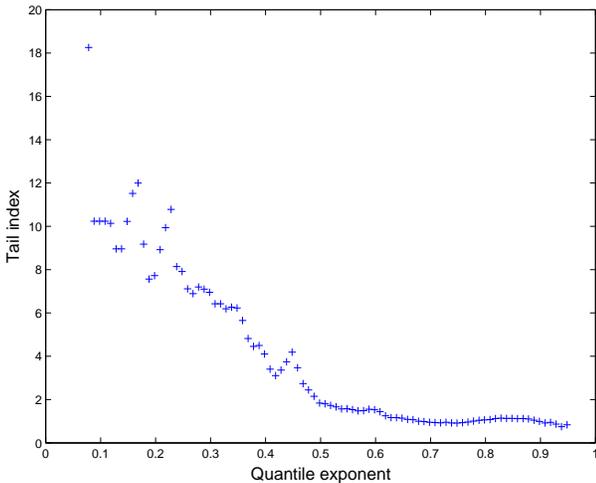

Figure 13: Hill plot, for a one hour segment in a busy period.

The Alternate Hill of a one hour extract (360 000 points) of the Drexel LAN traffic data are shown in Figures 13. The right point of the plot coincides with the 10% upper quantile, which accounts for nearly 80 % of the cumulated workload (see figure 4). A stability region can be found starting at the 2 % upper quantile with value of the tail index in the range $(1.5, 1.8)$: it is however difficult to determine a precise value (the Hill estimator in the range of 2 % to 5 % upper quantile is more like a monotonously declining line). This value may be considered as an evidence of heavy tail. Note how-ever that this tail index is not compatible with the values found for the extreme quantile (say above the 1 % upper quantile), where the tail index is much larger : this effect cannot only be attributed to the variance of the Hill plot but reflects the fact that the distribution above extreme quantiles are not well approximated by regularly varying tails because these models do not take into account the fact that the data is indeed bounded as shown on figure 3. As noted in the introduction however, figure 4 shows that modeling of upper quantiles is an important matter since these contribute significantly to the cumulative workload of the network.

6.2 Quantile-quantile regression plot (qqplot)

The quantile-quantile regression plot is also based on the property that, for a regularly-varying X_k with index α, the distribution of $X_{n-k+j,n}/X_{n-k,n}$, for large k, is approximately Pareto with shape parameter α.

If the data are Pareto distributed over $[1, \infty)$ with shape parameter α, then $\{(1 - j/(n + 1))^{-1/\alpha}, 1 \leq j \leq n\}$ are quantiles of their probability distribution whereas $\{X_{j,n}\}_{1 \leq j \leq n}$ are the corresponding quantiles of the empirical distribution function. This suggests to estimate α by plotting the empirical quantiles versus the theoretical quantiles (qq-plot) in log scales for both axis. Indeed, $\{(-\log(1-j/(n+1)), \log(X_{j,n})), 1 \leq j \leq n\}$ should be approximately a line with slope α^{-1}. Similarly as above, when X_k is regularly-varying with index α, and k is such that $k^{-1} + k/n \to 0$, then $X_{n-k+j,n}/X_{n-k,n}$, $1 \leq j \leq k$ is approximately Pareto with shape parameter α. The slope of regression $1 - \log(1 - j/(k + 1))$ through the points $\log(X_{n-k+j,n}/X_{n-k,n})$ can thus be used as an estimator of the tail index. This estimator is referred to as the *qq-plot* estimator. Although in the i.i.d. case its asymptotic variance is worse than the asymptotic variance of the Hill estimator, the variability of the qq-plot always seems to be less than that of the Hill estimator.

As in the case of the Hill plot, the *dynamic* qq-plot obtained by plotting $(k, \hat{\alpha}_{k,n})$, for $l < k < n$ is a useful tool. Another useful graph is the static qq-plot in which one represents the data $\log(X_{n-k+1,n}/X_{n-k,n})$ as a function of $\log(1-j/(k+1))$ together with the least-squares regression line. The slope of that line is used to compute the qq-estimator $\hat{\alpha}_{k,n}^{-1}$. interest to assess the validity of the
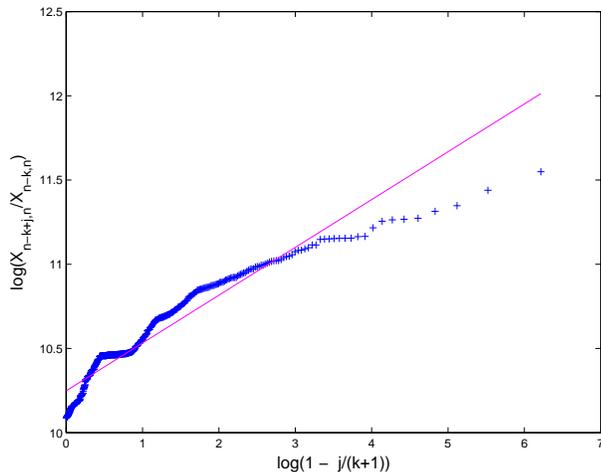
13

Figure 14: Static QQ-plot for a 1-hour segment in a busy period ($k = 500$).

regular tail variation model. In figure 14, the static qq-plot for a one hour ($n = 360000$) segment of the Drexel data (busy traffic period) is displayed for $k = 500$. On this example, the static qq-plot suggests that the upper-quantiles ($k = 500$ highest values out of $n = 360000$) are compatible with a regular variation model with tail index, as given by the slope of the regression line, about 3.52. This correspond to a distribution with very heavy tails (compared to those of a Poisson model for instance) but with finite variance.

7 Conclusions

This tutorial paper has discussed statistical models for analyzing long-range dependence and impulsive phenomena that appear in high-speed network traffic. The main challenge of this exciting rapidly evolving domain of engineering is that the characteristics of the data are so extreme that they have led to the development of new concepts and models, the most simple of which have been described in section 4. Finally, since we are indeed dealing with extreme behavior (very long term correlations in one case and very rare events in the other), the statistical performances of the estimation procedures become a very important concern which has been covered in some details in sections 5 and 6.

References

[1] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi and D. Veitch, "Wavelet and multiscale analysis for network traffic," this issue.

[2] J. Beran, *Statistics for Long-Memory Processes*. Chapman & Hall, New York, 1994.

[3] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger, "Long-range dendence in variable-bit-rate video traffic," *IEEE Trans. Comm*, vol. 43, no. 2, 1995.

[4] P.J. Brockwell and R.A. Davis, *Time series: Theory and methods*. Springer, 1991.

[5] M. E. Crovella and A. Bestravros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835-846, 1997.

[6] R. Davis and S. I. Resnick, "Limit theory for the sample covariance functions of moving averages," *Ann. Statist.*, vol. 44, pp. 533–558, 1986.

[7] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc SIG-COM*, London (UK), pp. 269-279, 1994.

[8] J. Geweke and S. Porter-Hudak, "The estimation and application of long memory time series models," *J. of Time Series Analysis*, vol. 4, pp. 221-238, 1983.

[9] L. Giraitis, P.M. Robinson, and A. Samarov, "Rate optimal semiparametric estimation of the memory parameter of the Gaussian time series with long range dependence," *J. of Time Series Analysis*, vol. 18, pp. 49-61, 1997.

[10] L. Giraitis, P.M. Robinson, and A. Samarov, "Adaptive rate optimal estimation of the long memory parameter," *J. Multivariate Analysis*, vol. 72, pp. 183-207, 2000.

[11] C.W.J. Granger and R. Joyeux. "An introduction to long memory time series and fractional differencing," *J. of Time Series Analysis*, vol. 1, pp. 15-30, 1980.

[12] M. Grossglauser and J. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Trans. Networking*, vol. 7, no. 5, Oct. 1999.

[13] C. A. Guerin, H. Nyberg, O. Perrin, S. I. Resnick, H. Rootzén and C. Starica, "Empirical testing of the infinite source Poisson data traffic model," Report 2000:4, Chalmers University of Technology, 2000.

[14] D. Heath, S. I. Resnick, and G. Samorodnitsky, "Heavy tails and long range dependence in on/off processes and associated fluid models", *Mathematics of Operation Research*, vol.23, pp. 145-165, 1998.

[15] M. Henry, "Robust automatic bandwidth for long-memory", *Journal of Time Series Analysis*, 22(3) pp. 293-317, 2001.

[16] D. P. Heyman and T. V. Lakshman, "What are the implications of long-rang dependence for VBR-video traffic engineering?" *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 301-317, 1996

[17] B. M. Hill, "A simple general approach to inference about the tail of a distribution", *Annals of Statistics*, vol.3, pp. 1163-1174, 1975.

[18] J. R. M. Hosking, "Fractional differencing", *Biometrika*, vol. 68, no. 1, pp. 165-176, 1981.

[19] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, pp. 770-808, 1951.

[20] C.M. Hurvich, E. Moulines, and P. Soulier, "The FEXP estimator for non Gaussian, potentially no stationary processes," Prépublication 133 de l'Université d'Evry Val d'Essonne, 2000.

[21] C. M. Hurvich and W. W. Chen, "An efficient taper for potentially overdifferenced long-memory time series," *J. of Time Series Analysis*, vol. 21, no. 2, pp. 155-180, 2000.

[22] J. Ilow, "Forecasting Network Traffic Using FARIMA Models with Heavy Tailed Innovations," in *Proc. ICASSP 2000*, Istanbul (Turkey), June 2000.

[23] A. Iouditsky, E. Moulines, and P. Soulier, "Adaptive estimation of the fractional differencing coefficient," *Bernoulli*, 7(5) p. 699-731, 2001

[24] A. Karasaridis, D. Hatzinakos, "Network heavy traffic modeling using alpha-stable self-similar processes," *IEEE Trans. Communications*, Vol. 49, no.7, pp. 1203-1214, July 2001.

[25] P. S. Kokoszka and M. S. Taqqu, "Can one use the Durbin-Levinson algorithm to generate infinite variance fractional ARIMA time series?" *J. of Time Series Analysis*, vol. 22, no. 3, pp. 317-338, 2001.

[26] P. S. Kokoszka and R. J. Bhansali, "Estimation of the long memory parameter: a review of recent developments and an extension," *Proceedings of the Symposium on Inference for Stochastic Processes*, I. V. Basawa, C. C.Heyde and R. L. Taylor, eds. IMS Lecture Notes, pp. 125-150, 2001.

[27] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACm Trans. Networking,* vol. 2, pp. 103-115, 1994.

[28] Z. Liu, N. Niclausse, C. Jalpa-Villanueva and S. Barbier, "Traffic Model and Performance Evaluation of Web Servers," technical report RR-3840, INRIA, 1999.

[29] S. B. Lowen and M. C. Teich, "Fractal renewal processes generate $1/f$ noise," *Phys. Rev. E*, vol. 46, no. 2, Feb 1993.

[30] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W. H. Freeman, San Francisco, 1982.

[31] T. Mikosch, S. I. Resnick, H. Rootzén, and A. Stegeman, "Is network traffic approximated by stable Lévy motion or fractional Brownian motion?" *Ann. Appl. Probab.*, to appear, 2001.

[32] T. Mikosch, S. I. Resnick, H. Rootzén, and A. Stegeman, "Is network traffic approximated by stable Lévy motion or fractional Brownian motion?," *Ann. Appl. Probab.*, 2001, to appear.

[33] E. Moulines and Ph. Soulier. "Broadband log-periodogram regression of time series with long-range dependence," *Annals of Statistics*, vol. 27, no. 4, pp. 1415-1439, 1999.

[34] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. Wiley, 1995.

[35] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *IEEE/ACM Trans Networking*, vol. 3, no. 3, June 1995.

[36] A.P. Petropulu, J-C. Pesquet, X. Yang, and J. Yin, "Power-law shot noise and relationship to long-memory processes," *IEEE Transactions on Signal Processing*, vol. 48, no. 7, July 2000.

[37] S. I. Resnick, "Heavy tail modeling and teletraffic data," *Annals of Statistics*, vol. 25, no. 5, pp. 1805-1848, 1997.

[38] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, 1994

[39] M.S. Taqqu, W. Willinger and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communication Review*, vol. 27, pp.5-23, 1997.

[40] M.S. Taqqu and V. Teverovsky, "On estimating the intensity of long-range dependence in finite and infinite variance time series" in *A practical guide to heavy tails: statistical techniques and applications*, R. Adler, R. Feldman and M.S. Taqqu editors, Birkaüser, Boston, 1998.

[41] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernets LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, Feb. 1997.

[42] W. Willinger and V. Paxson, Discussion of the paper by S. I. Resnick, *Annals of Statistics*, vol. 25, no. 5, pp. 1856-1866, 1997.

[43] X. Yang, A.P. Petropulu, "The extended alternating fractal renewal process for modeling traffic in high-speed communication networks," *IEEE Trans. Sig. Proc.*, vol. 49, no. 7, July 2001.