Heavy-Tailedness, Long-Range Dependence and Self-Similarity in Data Traffic

Sándor Molnár, Trang Dinh Dang , Attila Vidács

High Speed Networks Laboratory
Dept. of Telecommunications and Telematics
Technical University of Budapest
H–1117, Pázmány Péter sétány 1/D, Budapest, Hungary
Tel: (361) 463 3889, Fax: (361) 463 3107
E-mail: {molnar,trang,vidacs}@ttt-atm.ttt.bme.hu

Abstract

There is an increasing tendency in the traffic characterization of high speed networks to model the complex bursty traffic by fractal models exhibiting very appealing parsimony properties [9, 19]. This approach is supported by a number of traffic analysis studies based on several network measurements demonstrating different manifestations of self-similarity and long-range dependence [19].

In this paper a short framework of fractal traffic characterization is presented and our traffic analysis results concerning the relevance of fractal modeling of ATM WAN and Internet traffic are reported.

1 Introduction

Recent traffic measurement studies indicated that our traffic in modern communications networks (e.g. ATM, Internet) has *high variability* and *burstiness* over many time scales [19]. From a modeling point of view this phenomenon is difficult to characterize by Markovian models. An alternative but not very well elaborated type of modeling approach is to use the notion of *fractals* to characterize the complex "burst within burst" traffic structure.

The concept of fractal traffic modeling introduced the notion of *long-range dependence* and *self-similarity* to teletraffic theory. The first indication of the presence of fractal properties was published in [9] and since that time a number of similar results were reported investigating also the performance implications of fractal behaviour [10, 12, 18]. A good historical guide of traffic measurements, analysis and fractal traffic modeling can be found in [19].

A considerable amount of research is focused on finding the physical explanations of fractal behavior in network traffic. We do not have the final answer to this question but several important observations have been made. For example it was shown by Taqqu *et al.* [16] that the self-similar behavior of Ethernet LAN can be well explained by an ON/OFF model having heavy-tailed distributions with infinite variance of the ON and/or OFF periods. The aggregation of traffic generated by these ON/OFF sources produces self-similar traffic. Other results based on the $M/G/\infty$ queueing model were also published by Cox [3] and a more refined model by Kurtz [8] which provides explanations for self-similar traffic dynamics. A common and important observation of all these models is that *heavy-tailed* distributions play an important role in the observed fractal properties [20].

It should be noted that non-stationarities in network traffic can also produce properties detected by many statistical methods which are similar to fractal properties [4, 11]. Non-stationarity models can offer an alternative modeling approach to capture these properties [10].

In spite of the various publications in this field the framework of fractal traffic modeling (including multifractals, a recent development of this research field, see [6, 10] for details) is not well established. Especially, the connections between self-similarity, long-range dependence and heavy-tails are not clear in the present literature.

In this paper we provide a short summary of this framework and provide an analysis study of long-range dependence, self-similarity and heavy-tails of different traffic data including ATM and Internet. We have analyzed a large amount of traffic traces taken during a trial on the Swedish ATM Wide Area Network. The traces were measured by a custom-built measurement tool which is able to record more than 8 million consecutive cell arrivals. The analysis of heavy-tailed distributions requires enormous sized data which we collected exploiting this special ability of our recording instrument. Thus we had the opportunity to get results supported by well established statistics. Besides ATM data we also performed a comprehensive analysis on Internet data, the results of which are reported in this paper.

We give a short overview about the mathematical basics of heavy-tailed distributions. For testing heavy-tailed distributions we used different statistical methods such as variants of QQ-plot, Hill-method and the De Haan's moment method. Similarly, for the statistical analysis of long range dependence and self-similarity a number of methods have been used (variance-time plot, R/S plot, periodogram plot, Whittle estimator). The results are presented with explanations.

2 The Concepts of Heavy-Tails, Long-Range Dependence and Self-Similarity

In this chapter a brief overview about the fractal traffic framework is given. The important definitions and only the most important properties are mentioned here. For details see the indicated references.

2.1 Heavy-tailed distributions

Let X be a non-negative random variable with distribution function F.

Definition 1 *[13] F is said to be heavy-tailed if*

$$1 - F(x) = x^{-\alpha} L(x), \tag{1}$$

where L is slowly varying at ∞, i.e., $\lim_{x\to\infty} L(tx)/L(x) = 1$, $t > 0$.

For example, in the simplest case $L(x) \equiv 1$, and the distribution with $F(x) = 1 - x^{-\alpha}$ is the so-called Pareto distribution. In the general case this definition can be reformulated by using the concept of regular varying functions [13]:

Definition 1′ *F is heavy-tailed if and only if $1 - F$ is regularly varying with index $-\alpha$, i.e., $\lim_{t\to\infty}(1 - F(tx))/(1 - F(t)) = x^{-\alpha}$, $x > 0$.*

For heavy-tailed distributions tails decay like a power, hence they are also called *power law* or *hyperbolic* distributions. (This behaviour is in contrast to the exponentially decaying tails of *light tailed* distributions, where $1 - F(x) \sim e^{-cx}$, $c > 0$, as this is the case for exponential distributions, for example.) Heavy-tailed distributions are also *subexponential in the wide sense* (see [5] for the exact mathematical definition) in the sense that the rate of decay is slower than exponential. Similarly, the term *long tailed* (see [7] for details) can also be used for heavy-tailed distributions. (It should be noted, however, that these terms cannot be treated as equivalent definitions. The set of heavy-tailed distributions is the subset of both the subexponential distributions and the long-tailed distributions.)

The following properties can be derived from heavy-tailedness. When $X \geq 0$ has a heavy-tailed distribution, a simple condition for the existence of the moments can be given [13],

$$E(X^\beta) < \infty, \quad \beta < \alpha, \tag{2}$$
$$E(X^\beta) = \infty, \quad \beta \geq \alpha. \tag{3}$$

For example, if $1 \leq \alpha < 2$, F has finite mean but infinite variance. This phenomenon is known in the literature as the *Noah effect* or *infinite variance syndrome* .

Suppose now that $X_1, \ldots, X_n$ are iid samples with heavy-tailed distribution. Denote the partial sum of $X_1, \ldots, X_n$ by $S_n = X_1 + \cdots + X_n$ and their maximum by $M_n = \max(X_1, \ldots, X_n)$. Then heavy-tailedness implies $P(S_n > x) \sim P(M_n > x)$, as $x \to \infty$ [5]. (Note, that the distribution family satisfying the above property is called *subexponential* [5]. For example, the Weibullian distribution is subexponential.) This striking feature reveals the fact that in case of heavy-tails the large samples dominate since the probability of 'being large' is non-negligible.

2.2 Long-range dependence

Let X_t be a stationary process with autocorrelation function $\rho(\cdot)$ and power spectral density $f(\cdot)$.

Definition 2 *[1] X_t is called a stationary process with long range dependence (LRD or long memory) if there exists a real number $H \in (0.5, 1)$ and a constant $c_\rho > 0$ such that*

$$\lim_{k\to\infty} \frac{\rho(k)}{c_\rho k^{2H-2}} = 1, \tag{4}$$

where H is called the Hurst parameter and measures the degree of LRD.

The following statement is equivalent to the definition: if X_t is LRD, then there exists a constant $c_f > 0$ such that

$$\lim_{\nu\to 0} \frac{f(\nu)}{c_f |\nu|^{1-2H}} = 1. \tag{5}$$

LRD is also referred to as the *Joseph effect or the* persistence phenomenon. In this case $\sum_k \rho(k) = \infty$. (Note, that the non-summability of the autocorrelation function is not equivalent but more general property than the definition used.) In contrast, for *short range dependent (SRD or short memory) processes the autocorrelation function is geometrically bounded, i.e.,* $\lim_{k\to\infty}\rho(k)/c^k = 1$, $0 < c < 1$ *and thus* $\sum_k |\rho(k)| < \infty$. *Processes for which Eq.(4) holds with* $H < 0.5$ *the* $\sum_k |\rho(k)| < \infty$. *These processes are called* intermediate memory processes [2].

Examples for LRD processes are fractional Gaussian noise and F-ARIMA processes, while all Markovian, ARMA and finite memory processes are short range dependent.

2.3 Self-similarity

Definition 3 *[14] The real-valued process* $\{Y(t), t \in \mathbf{R}\}$ *is self-similar with index* $H > 0$ *(H-ss) if for all* $a > 0$, *the finite-dimensional distributions of* $\{Y(at)\}$ *are identical to the finite-dimensional distributions of* $\{a^H Y(t)\}$; *i.e., if for any* $d \geq 1$, $t_1, t_2, \ldots, t_d \in \mathbf{R}$ *and any* $a > 0$,

$$(Y(at_1), Y(at_2), \ldots, Y(at_d)) \stackrel{d}{=} \left(a^H Y(t_1), a^H Y(t_2), \ldots, a^H Y(t_d)\right). \quad (6)$$

A non-degenerate H-ss process cannot be stationary, but can have stationary increments.

Definition 4 *[14] The process* $\{Y(t), t \in \mathbf{R}\}$ *is called* H*-sssi if it is self-similar with index* H *and has stationary increments.*

If $\{Y(t)\}$ is a (non-degenerate) H-sssi finite variance process, then $0 < H \leq 1$. The increment sequence of $\{Y(t)\}$ in discrete time can be defined as $X_k = Y(k) - Y(k-1)$, $k = 1, 2, \ldots$. Define the m-aggregated time series $X^{(m)}$ and its autocorrelation function $r^{(m)}(\cdot)$ as follows:

$$X_k^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad (7)$$

$$r^{(m)}(k) = E X_k^{(m)} X_0^{(m)}. \quad (8)$$

The interesting range of H is $0.5 < H < 1$ for traffic modeling because H-sssi $Y(t)$ processes with $H < 0$ are not measurable and represent pathological cases while for the $H > 1$ case the autocorrelation of the incremental process does not exist. The range of $0 < H < 0.5$ can also be excluded from our practice because in this case the incremental process is SRD. For practical purposes the range of $0.5 < H < 1$ is only important. In this range the autocorrelation of the incremental process is $r(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H})]$. This incremental process is LRD which shows the connection between self-similar and long-range dependent processes.

For an exactly (second-order) self-similar process

$$\begin{aligned} \operatorname{var}(X^{(m)}) &= \frac{1}{m^{2-2H}} \operatorname{var}(X), && (9) \\ r^{(m)}(k) &= r(k). && (10) \end{aligned}$$

A weaker condition is the following: A process X is said to be asymptotically (second-order) self-similar if for all k large enough

$$\lim_{m\to\infty} r^{(m)}(k) = r(k). \quad (11)$$

The only Gaussian process that is self-similar and has stationary increments is called fractional Brownian motion (FBM) and its increment process is referred to as fractional Gaussian noise (FGN).

3 Traffic Measurements

During our study, some real data sets were analyzed. These data sets were measured at different environments in ATM networks and on the Internet. The ATM measurements were performed at Telia Research on the SUNET WAN ATM network. The Internet data bases are freely available from the Internet Traffic Archives [17]. This section describes in detail these traffic measurements.

3.1 SUNET ATM networks

The configuration of the measurement is shown in Figure 1. As a business customer of Telia, the Swedish network operator, different parts of the Swedish University Network (SUNET) are attached to Telia's ATM wide area networks. During summer 1996, the aggregated traffic on a SUNET LAN interconnection was analyzed in the framework of a common trial between the SUNET community and Telia research. The LAN traffic of universities in the northern region, around Uppsala is connected to an FDDI backbone, which is further connected on R1, R2 routers and a 34 Mbps

PDH link to the ATM backbone in Stockholm. This network joins the northern LANs of SUNET to the international Internet backbone and to the southern university networks around Göteborg. A CBR (Constant Bit Rate) connection with 38.16 Mbps cell rate was established on the SDH link between the routers R4 and R5 for the trial. The measurements reported here were performed on the connection between Uppsala and Göteborg. ATM traffic streams were duplicated by means of optical splitters avoiding impacts on original traffic flows. The duplicated traffic streams were routed on dedicated links to Telia Research in Haninge, where almost one hundred traffic traces were collected with more than 8 millions cell arrivals in each trace.

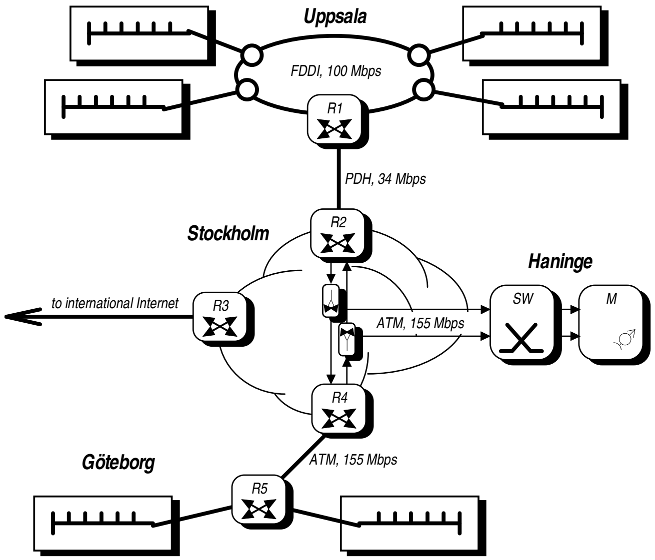

Figure 1: The configuration of the SUNET measurements

This traffic was an ordinary mix of common Internet traffic types such as HTTP, FTP, Telnet, Chat, etc. and can be considered as a typical sample from today's traffic with integrated services and applications. From the trace, the process of ATM cell arrival counts in consecutive time-windows of 400 cell time was used for analysis.

3.2 IP traffic traces

This trace is the result of an hour long Ethernet measurement run from 14:00 to 15:00 on Friday, January 21, 1994. The tracing was done on the Ethernet DMZ network which provided all the incoming and outgoing traffic of the Lawrence Berkeley Laboratory, located in Berkeley, California. The raw traces were made using `tcpdump` on a Sun SparcStation using the BPF kernel packet filter.

The measurement captured arrival timestamps in microsecond precision of TCP, UDP, TCP SYN/FIN/RST, encapsulated IP and other IP packets in five files, respectively. After processing these files, a set of around $300,000$ IP packet arrivals in consecutive time-windows, equally $0.021 sec$, was selected for analysis.

3.3 WWW traffic traces

These measurements were done at Boston University's Computer Science Department. In order to capture all of the Web activity on a Local Area Network (LAN), researchers modified the Web browser NCSA Mosaic and installed it for general use. After that Mosaic browsers could write down all working activities of browsers in a *log* file. Each line in a *log* corresponds to a single URL requested by the user; it contains the machine name, the timestamp when the request was made, the user id number, the URL, the size of the document (including the overhead of the protocol) and the object retrieval time in seconds (reflecting only actual communication time, and not including the intermediate processing performed by Mosaic in a multi-connection transfer). These traces contain records of the HTTP requests and user behavior of a set of Mosaic clients running in a general computing environment at the department. This environment consists principally of 37 SparcStations 2 workstations connected in a local network, which is divided into two sub-nets. Each workstation has its own local disk; *log*s were written to the local disk and subsequently transferred to a central repository. The data collection then took about 5 months from 17 January 1995 until 8 May 1995.

In this study we consider only the characteristics of the file sizes transmitted over the Internet. So a small C routine was implemented to extract this information from over $6,000$ *log* files. Around $230,000$ unique file sizes were recorded. As the suggestion of some previous studies, this data set—called the Web file sizes data set or the WFS set—may contain heavy-tailed properties.

4 Analysis

There are number of methods developed in statistics to investigate the presence of heavy-tailedness, long-range dependence and self-similarity [1, 5, 13, 15, 20]. However, working in practice a number of difficulties arises. A practical problem is that in order to have a reliable description of these properties a huge data set is needed. Moreover, most of our statistical tools require different assumptions about the data which can be very difficult to check (e.g. stationarity). We have to use more statistical tools to check different manifestations of the investigated property to avoid pitfalls and mis-interpretations.

4.1 Testing for heavy-tails

For estimating the index α of the heavy-tailed distributions, a lot of estimating methods can be found in the literature. Among them the modified QQ-plot, the Hill estimator and DeHaan's moment estimator seem to be the better known ones.

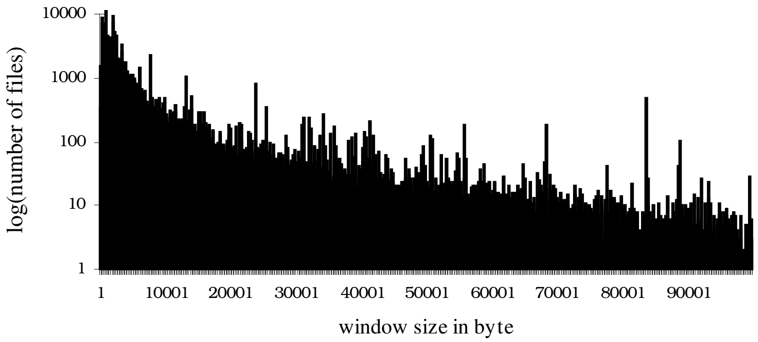

Figure 2: The histogram of the WWW file sizes set on the log-lin graph

These methods should be used to estimate the index α when there is a certain evidence which indicates the possible existence of heavy-tail. The log-linear plot of empirical histogram is a good tool for visualizing the tail behaviour. Figure 2 is the histogram built from the WFS set. To draw the histogram the number of files whose size in bytes is between 1 and 100, then 101 and 200, and so on, was counted and displayed on a log scale against the window size. Note that in this scale the histogram of a light-tailed process should be a straight line. In our case, the slow decay observed on the figure shows that WFS data set may be heavy-tailed.

4.1.1 Modified QQ-plot

The modified QQ-plot is based on the QQ-plot, which is a widely known testing method in statistics. Modified QQ-plot is adapted to the problem of detecting heavy-tails and for estimating index α. The description of this method is also discussed in [5].

Suppose $\{X_1, X_2, \ldots, X_n\}$ are iid random samples of distribution F. Pick k upper order statistics $X_1^* \geq X_2^* \geq \ldots \geq X_k^* = u$ and neglect the rest. The plot of

$$\left\{ \left(\log X_j^* - \log u, \quad -\log\left(\frac{j}{k+1}\right)\right), \quad 1 \leq j \leq k \right\} \tag{12}$$

should roughly look like a straight line with slope $= \alpha$, if data is approximately Pareto or even if $1 - F$ is regularly varying (see Definition 1').

The main idea of using modified QQ-plot follows the assumption: if $X_1^* \geq X_2^* \geq \ldots \geq X_k^*$ are samples from a distribution F and k is large enough, the distribution function F at $x = X_j^*$ can be estimated by

$$P(x < X_j^*) = F(X_j^*) \approx 1 - \frac{j}{k+1}. \tag{13}$$

Figure 3 is the modified QQ-plot of WFS data set. It can be seen in the figure that the plot is not exactly a straight line but a regression line can be fitted to the points with a small deviation. The slope provides the estimate of α to be 0.73.

4.1.2 Hill estimator

The Hill estimator is a statistical test which is optimized for estimating the index α for distributions close to Pareto [13].

Suppose that $X_1, X_2, \ldots, X_n$ are iid samples from a distribution F. Let $X_1^* \geq X_2^* \geq \ldots \geq X_n^*$ be the order statistics. If F is a Pareto distribution

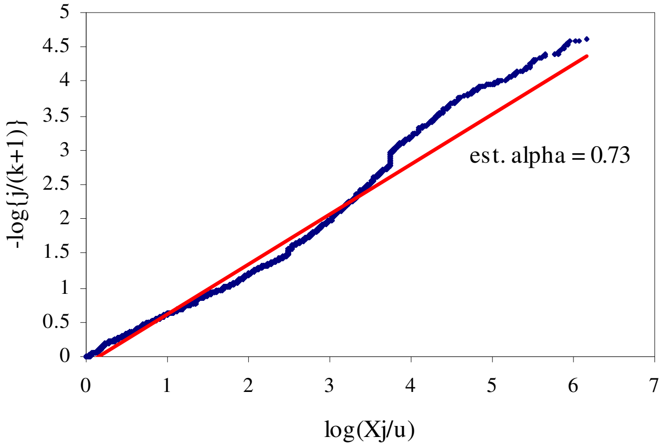

Figure 3: The modified QQ-plot of the WWW file sizes set

then the Hill estimate can be taken from the maximum likelihood estimator (MLE) of $\log X_1^*, \log X_2^*, \ldots, \log X_k^*$

$$H_{k,n} = \hat{\alpha}^{-1} = \frac{1}{k}\sum_{j=1}^{n} \log X_j^* - \log X_k^* \tag{14}$$

where k is the number of upper-order statistics used in the estimation. Thus the Hill estimate of index α is

$$\hat{\alpha} = \frac{1}{H_{k,n}}. \tag{15}$$

As discussed before, the Hill estimator is designed for Pareto distributions, so it can be misleading when dealing with some heavy-tailed data sets, which are not exactly Pareto.

The Hill estimation of WFS data set can be seen on Figure 4. The plot goes fast to its stable value 0.67. It is the estimate of index α of the WFS distribution tail.

4.1.3 De Haan's moment method

De Haan's moment estimator (see also [13]) is designed to estimate parameter γ from random samples in the domain of attraction of the extreme

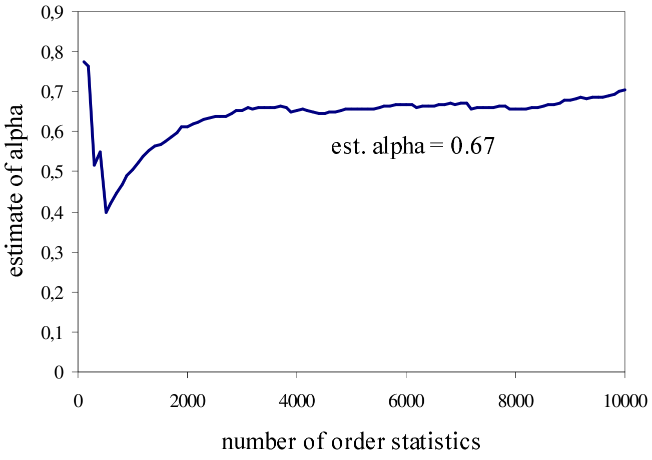

Figure 4: The Hill plot of the WWW file sizes data set

value distributions G_γ.

$$G_\gamma = e^{-(1+\gamma x)^{-1/\gamma}}, \qquad \gamma \in R, \quad 1 + \gamma x > 0 \tag{16}$$

When the estimate of γ is positive, it also provides the estimate of index α with $\hat{\alpha} = 1/\hat{\gamma}$. This method also provides another method of deciding whether a distribution is heavy-tailed or not. If the estimate of γ is negative or very close to zero, it suggests that the sample distribution does not possess heavy-tailedness.

De Haan's moment estimator is defined as follows: Let $X_1^* \geq X_2^* \geq \ldots \geq X_n^*$ be the order statistics from a random sample of size n. Define for $r = 1, 2$ and for k upper-order statistics

$$H_{k,n}^{(r)} = \frac{1}{k}\sum_{i=1}^{k}\left(\log \frac{X_i^*}{X_{k+1}^*}\right)^r. \tag{17}$$

De Haan's estimate of γ can be calculated by the form

$$\hat{\gamma} = H_{k,n}^{(1)} + 1 - \frac{1}{2\left(1 - \frac{(H_{k,n}^{(1)})^2}{H_{k,n}^{(2)}}\right)}. \tag{18}$$

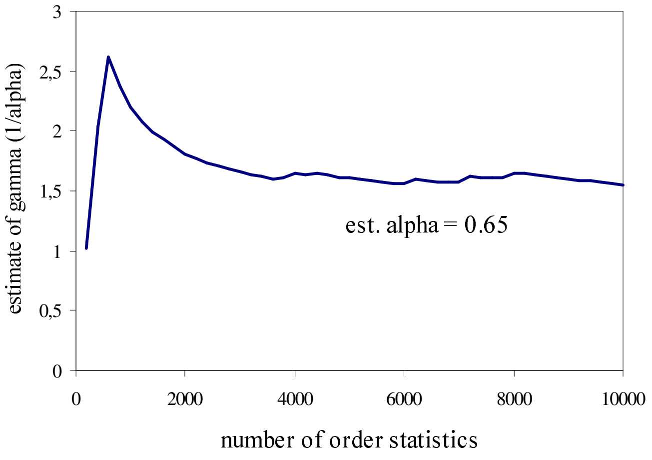

Figure 5: The DeHaan estimation of the WWW file sizes set

Figure 5 shows the plot result generated by De Haan's testing method. The estimate of α in this case, 0.65, is a bit smaller than in the Hill's case. It may be the effect of the smoothing technique used in De Haan's algorithm.

From the results presented above we can conclude that the file sizes transmitted over the Internet (WFS set) have heavy-tails and can be modeled by a Pareto distribution with parameter α to be approximately 0.7.

4.2 Testing for long-range dependence

The phenomenon of long-range dependence in data traffic has been a hot topic in traffic modeling in the recent several years. When trying to estimate the Hurst parameter most authors use the following methods: variance-time plot, R/S plot, periodogram, and Whittle estimator. These testing methods are discussed in detail in [1] and [15], for example.

Using LRD tests and other statistical tests, it is difficult to make reliable conclusions about the self-similarity of traffic. Note that in most cases statistical methods cannot prove whether an empirical data set is taken from an exactly self-similar process. Instead, as discussed in subsection 2.3, a data set may only have the property of second-order or asymptotically second-order self-similarity.

4.2.1 Variance-time plot

Based on property Eq.(9) of a LRD process, the variance-time plot is defined as follows: Using the property of LRD given in Eq.(9) with $\beta = 2 - 2H$ we have

$$\log\left(\mathrm{var}(X^{(m)})\right) = \log\left(\mathrm{var}(X)\right) - \beta \log(m). \qquad (19)$$

Because $\log(\mathrm{var}(X))$ is a constant independent of m, if we plot $\mathrm{var}(X^{(m)})$ versus m on a log-log graph, the result should be a straight line with a slope of $-\beta$. The Hurst parameter can be calculated from β by the formula $H = 1 - (\beta/2)$. The plot can be easily generated from the data series X by generating the aggregated processes of X at different levels of m and then computing their empirical variance. A plot with slope values between -1 and 0 suggests LRD.

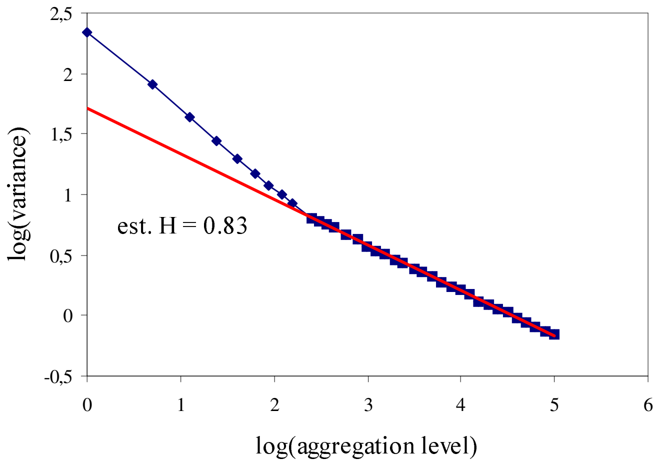

Figure 6: The variance-time plot of the IP data set

The variance-time plot of IP data set is drawn on Figure 6. It is surprising that there is a breaking point in the picture. From a certain large value of the time unit, the slope takes up a bigger value. Anyway, by the Definition 2 of LRD it is an asymptotic characteristics, so the Hurst parameter should be estimated by the slope on the higher aggregation levels. The estimate of H was 0.83.

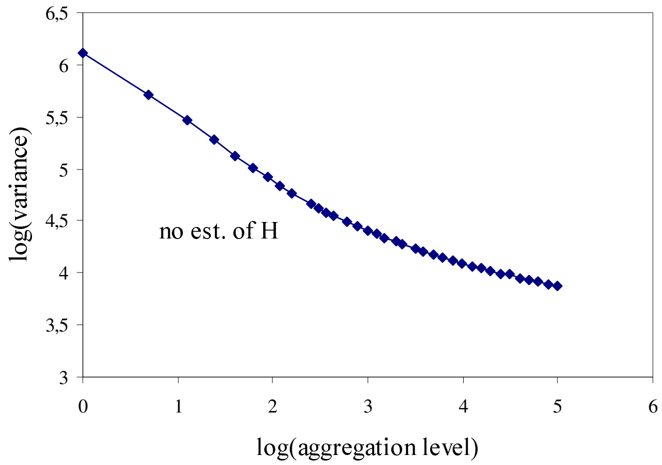

Figure 7: The variance-time plot of the SUNET data set

In the case of SUNET data (Figure 7), the variance-time plot clearly looks like a curve rather than a straight line. From this result the estimation of Hurst parameter is not possible.

4.2.2 R/S plot

For a stochastic process X defined in discrete time $\{X_j : j = 1, 2, \ldots, n\}$, the rescaled range of X over a time interval n is defined as the ratio R/S:

$$\frac{R}{S} = \frac{\max\{W_i : i = 1, 2, \ldots, n\} - \min\{W_i : i = 1, 2, \ldots, n\}}{\sqrt{\mathrm{var}(X)}} \tag{20}$$

where $W_i = \sum_{k=1}^{i}(X_k - \bar{X}), i = 1, 2, \ldots, n$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It can be proven for any stationary process with LRD that the ratio R/S has the following characteristics for large n:

$$\frac{R}{S} \approx \left(\frac{n}{2}\right)^{H} \tag{21}$$

which is known under the name *Hurst effect*. Thus if we plot R/S versus n on a log-log graph $\log(R/S) \approx H\log(n) - H\log 2$, the plot should fit a straight line with slope H.

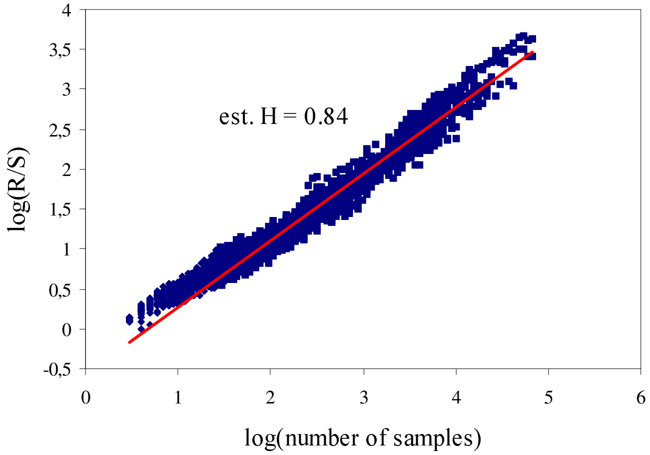

Figure 8: The R/S estimation of H of the IP data set

Using this algorithm, the R/S analysis of IP data set was provided and can be seen in Figure 8. Data points are scattered around a straight line, which means that IP packet arrivals seem to be LRD with Hurst parameter $H = 0.84$, which is the estimate from the slope of regression line.

Figure 9 is the R/S plot of the SUNET data set. The plot also displays a break point. The LRD parameter of the SUNET set looks not to be the same over all scale values. The estimate of H in this case, by Definition 2, should be calculated at the bigger values of k (number of samples), where it equals 0.95 (!).

4.2.3 Periodogram

This testing method is based on property Eq.(5) of LRD processes, namely, the power spectral density of a LRD process obeys a power law near the origin. So

$$\log f(\nu) \sim -\gamma \log \nu, \qquad \text{as} \quad \nu \to 0, \tag{22}$$

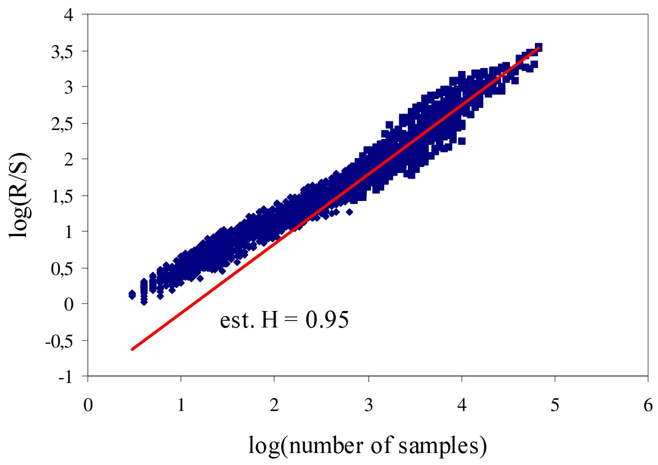

Figure 9: The R/S estimation of H of the SUNET data set

where $\gamma = 2H-1$. The spectral density of a discrete-time stochastic process is defined as

$$f(\nu) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} r(k) e^{ik\nu} \tag{23}$$

where σ^2 is the variance and $r(k)$ is the autocorrelation function.

Since the spectral density is the Fourier transform of the autocorrelation function, an estimate of the spectral density can be obtained by doing the Fourier transform on the estimate of the autocorrelation function. (This in fact produces a good estimate under certain reasonable conditions.) This estimator is referred to as a periodogram, and is defined as

$$I(\nu) = \frac{1}{2\pi n} \left| \sum_{k=1}^{n} (X_k - \bar{X}) e^{ik\nu} \right|^2 . \tag{24}$$

The periodogram plot is the graph of $\{\log \nu_j, \log I(\nu_j)\}$, $j = 1, 2, \ldots, M$ where $\nu_j = 2\pi j/n$ and M is always chosen to be $n/4, n/8, n/16$ or $n/32$ depending on how large n is. Following Eq.(22), the plot should be a straight line with slope $-\gamma = 1 - 2H$.

The periodogram plot of IP data set is shown on Figure 10. The estimate of H in this case is 0.82.

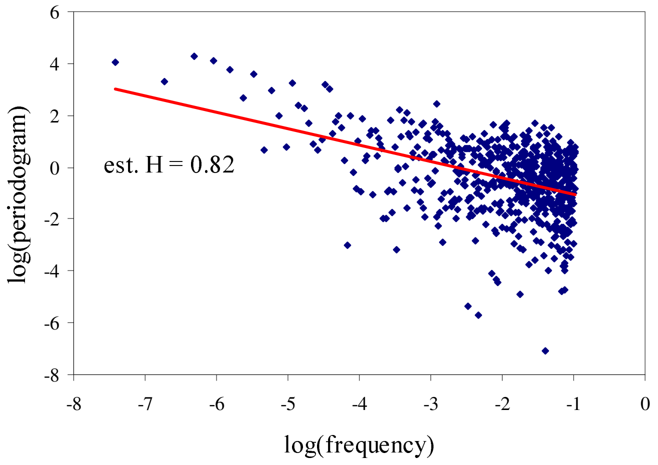

Figure 10: The periodogram estimation of the IP data set

4.2.4 Whittle estimator

The Whittle estimator is a concrete application of the maximum likelihood method (MLE). On the other hand, the Whittle estimation is based on the periodogram. So in most cases these methods provide the same estimates of the Hurst parameter.

The Whittle estimator was suggested to estimate the Hurst parameter of Fractional Gaussian Noise (FGN), which is an exactly self-similar process. If the data is from a FGN process, the estimate of H is the value that minimizes the function $Q(H)$:

$$Q(H) = \int_{-\pi}^{\pi} \frac{I(\nu)}{f(\nu, H)} \, \mathrm{d}x + \int_{-\pi}^{\pi} \log f(\nu, H) \, \mathrm{d}\nu. \tag{25}$$

To calculate the value of $Q(H)$, we consider the behavior of the spectral density of the process close to the origin (see Eq. (5))

$$f(\nu, H) \approx c_f |\nu|^{1-2H}, \qquad \nu \to 0 \tag{26}$$

Remember that the power spectral density function $f(\nu, H)$ can be esti-

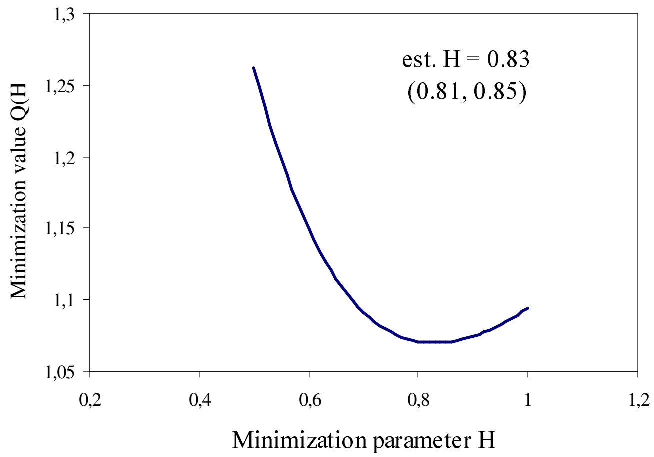

Figure 11: The Whittle estimation of the IP data set

mated by the periodogram $I(\nu)$, so the estimate of c_f is

$$\hat{c_f} = \hat{c_f}(H) = \frac{1}{M}\sum_{j=1}^{M}\frac{I(\nu_j)}{\nu_j^{1-2H}}. \tag{27}$$

So in the discrete case

$$Q(H) = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{I(\nu_j)}{\hat{c_f}(H)\,\nu_j^{1-2H}}\right) + \log\left(\hat{c_f}(H)\,\nu_j^{1-2H}\right) \tag{28}$$

where $\nu_j = 2\pi j/n$, frequencies are summed up to $2\pi M/n$ and M is always chosen to be $n/4$, $n/8$, $n/16$ or $n/32$ depending on how large n is. Inserting Eq.(27) into Eq.(28), we have

$$Q(H) = 1 + \log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{I(\nu_j)}{\nu_j^{1-2H}}\right) - (2H-1)\frac{1}{M}\sum_{j=1}^{M}\log(\nu_j). \tag{29}$$

The Whittle estimator also provides the confidence interval (95%) of H at 1.96σ, where

$$\sigma^2 = \frac{4\pi}{n}\,Q(H). \tag{30}$$

The disadvantages of this method are that we need to know the parametric form of the spectral density of the process and it takes a lot of time to calculate the result.

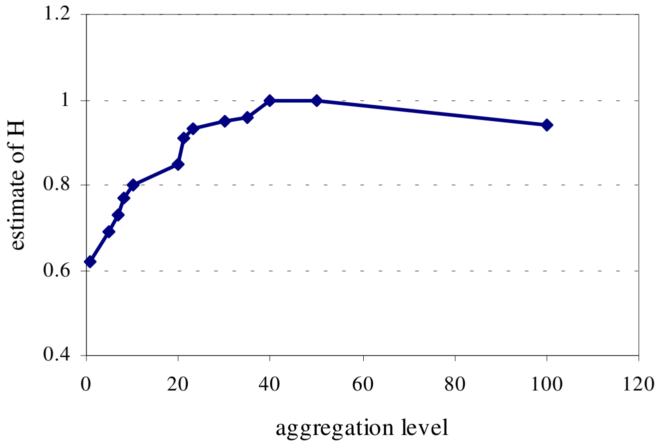

Figure 12: The periodogram and Whittle estimation of the SUNET data set

Figure 11 shows the results of the Whittle estimation of IP data set. The result is 0.83 with 95% confidence interval (0.81, 0.85). All the four statistical methods of the IP data set support our final conclusion that this IP traffic is LRD with Hurst parameter about $H = 0.83$. We found that this traffic has similar LRD structure for more investigated aggregation levels so it seems to be consistent with asymptotically second-order self-similarity.

On the contrary, the estimate of H of the SUNET data set provided by the periodogram plot and Whittle estimator is quite different from the IP data set. The values of H change with the values of aggregation levels. This result is plotted in Figure 12. As we demonstrated above the variance-time plot and R/S estimation provided non-evaluable results for this data set. All of these results suggest that the SUNET traffic is not consistent with self-similarity but has a more complex structure. In Figure 12 we can see that a different parameter estimate is obtained for different aggregation levels. This indicates that the scaling parameter is not constant at all time-scales but changes as we alter the time scale. This observation suggests that this traffic has no self-similar but rather multifractal structure. The detailed analysis of the multifractality of the SUNET data is one of our future research topics.

5 Conclusion

We presented a brief overview of the framework of fractal traffic characterization with the important mathematical concepts including heavy-tails, long-range dependence and self-similarity. We chose teletraffic data taken from both a living ATM WAN and the Internet to determine whether or not these data sets are consistent with heavy-tailed, long-range dependence and self-similarity properties. For the statistical analysis we used different methods to reveal these properties.

Our results demonstrate that the file sizes transmitted over the Internet (WFS set) have heavy-tails and can be modeled by a Pareto distribution with parameter α to be approximately 0.7.

We also concluded that the IP traffic is LRD with Hurst parameter about $H = 0.83$ and seems to be asymptotically second-order self-similar.

On the contrary, we have found that the investigated ATM WAN traffic data is not consistent with self-similarity and we detected the presence of a possible multifractal structure. Our future research will address the detailed investigation of this conjecture.

Acknowledgment
The reported ATM WAN measurements were carried out by the CARAT laboratory of Telia Research in Sweden. We would like to thank Nils Björkman, Urban Hansson, Alexander Latour-Henner and Aziz Miah for the measurements.

References

[1] J. Beran. *Statistics for Long-Memory Processes.* Chapman & Hall, One Penn Plaza, New York, NY 10119, 1994.

[2] P. J. Brockwell, R. A. Davis *Time Series: Theory and Methods.* Spinger–Verlag, 1986.

[3] D. R. Cox. *Statistics: An appraisal*, Long-range dependence: a review, pp. 55–74. Iowa State University Press, 1984.

[4] N. G. Duffield, J. T. Lewis, and N. O'Connell. "Statistical issues raised by the Bellcore data". 11th Teletraffic Symposium, Cambridge, March 1994.

[5] P. Embrechts, C. Klüppelberg, and T. Mikosh. *Modeling Extremal Events for Insurance and Finance.* Springer-Verlag, Berlin Heidelberg, 1997.

[6] A. Feldman, A. C. Gilbert and W. Willinger. "Data Networks as Cascades: Investigating the Multifractal Nature of Internet WAN Traffic". Computer Communication Review. Vol. 28. No. 4. pp. 42–55, 1998.

[7] P.R. Jelenković and A.A. Lazar. "Asymptotic results for multiplexing subexponential on-off sources". Submitted for publication to Advances in Applied Probability. To appear in part in: INFOCOM'97 Kobe, Japan, April 1997; ITC 15, Washington, D.C., USA, June 1997; INFORMS Applied Probability Conference, June 30-July 2, 1997, Boston (invited talk)., July 1996.

[8] T. G. Kurtz. *Stochastic Networks: Theory and Applications*, Limit theorems for workload input models, pp. 339–366, Clarendon Press, Oxford, UK, 1996.

[9] W. E. Leland, M. S. Taqqu, W. Willinger, and D. W. Wilson. "On the self-similar nature of Ethernet traffic" (extended version). IEEE/ACM Transactions on Networking, 2(1):1-15, February 1994.

[10] S. Molnár, I. Maricza (eds.) "Source Characterization in Broadband Networks". COST 257 Interim Report, January 1999.

[11] S. Molnár, A. Vidács, A. A. Nilsson. "Bottlenecks on the Way Towards Fractal Characterization of Network Traffic: Estimation and Interpretation of the Hurst Parameter". International Conference on the Performance and Management of Communication Networks (PMCCN'97), Tsukuba, Japan, 17-21 November 1997.

[12] S. Molnár and A. Vidács. "On modeling and shaping self-similar ATM traffic". 15th International Teletraffic Congress, Washington, D.C., USA, June 1997.

[13] S.I. Resnick. Heavy tail modeling and teletraffic data. *The Annals of Statistics*, 25(5):1805–1869, 1997.

[14] G. Samorodnitsky and M.S. Taqqu. *Stable Non-Gaussian Random Processes*. Chapman & Hall, One Penn Plaza, New York, NY 10119, 1994.

[15] M. S. Taqqu, Vadim Teverovsky, and Walter Willinger. "Estimators for long-range dependence: an empirical study". *Preprint*, 1995.

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

[17] The Internet Traffic Archive. Http://ita.ee.lbl.gov/index.html.

[18] A. Vidács, S. Molnár, G. Gordos, I. Cselényi. "The impact of long-range dependence on cell loss in an ATM wide area network". GLOBECOM'98, Sydney, Australia, November 1998.

[19] W. Willinger, M. S. Taqqu, and A. Erramilli. "A bibliographical guide to self-similar traffic and performance modeling for high speed networks: Stochastics Networks, Theory and Applications". 339-366, Oxford University Press, 1996.

[20] W. Willinger, V. Paxson, M. S. Taqqu. Self-similarity and heavy-tailes: structural modeling of network traffic. In J. Adler, R. E. Feldman, and M. S. Taqqu, editors, *A practical guide to heavy tails: statistical techniques and applications*, Birkhäuser, 1998.