10-1-2007

Towards Finding Efficient Tools for Measuring the Tail Index and Intensity of Long-range Dependent Network Traffic

Karim Mohammed Rezaul

Vic Grout
Glyndwr University, v.grout@glyndwr.ac.uk

Towards Finding Efficient Tools for Measuring the Tail Index and Intensity of Long-range Dependent Network Traffic

Abstract
Many researchers have discussed the effects of heavy-tailedness in network traffic patterns and shown that Internet traffic flows exhibit characteristics of self-similarity that can be explained by the heavy-tailedness of the various distributions involved. Self-similarity and heavy-tailedness are of great importance for network capacity planning purposes in which researchers are interested in developing analytical methods for analysing traffic characteristics. Designers of computing and telecommunication systems are increasingly interested in employing heavy-tailed distributions to generate workloads for use in simulation - although simulations employing such workloads may show unusual characteristics. Congested Internet situations, where TCP/IP buffers start to fill, show long-range dependent (LRD) self-similar chaotic behaviour. Such chaotic behaviour has been found to be present in Internet traffic by many researchers. In this context, the 'Hurst exponent', H, is used as a measure of the degree of long-range dependence. Having a reliable estimator can yield a good insight into traffic behaviour and may eventually lead to improved traffic engineering. In this paper, we describe some of the most useful mechanisms for estimating the tail index of Internet traffic, particularly for distributions having the power law observed in different contexts, and also the performance of the estimators for measuring the intensity of LRD traffic in terms of their accuracy and reliability.

Keywords
Keywords- Self-similarity, LRD, Tail Index, Hurst parameter

Towards Finding Efficient Tools for Measuring the Tail Index and Intensity of Long-range Dependent Network Traffic

Karim Mohammed Rezaul and Vic Grout
Centre for Applied Internet Research (CAIR), University of Wales, NEWI, Wrexham, UK
{karim|vic}@cair-uk.org

Abstract **-Many researchers have discussed the effects of heavy-tailedness in network traffic patterns and shown that Internet traffic flows exhibit characteristics of self-similarity that can be explained by the heavy-tailedness of the various distributions involved. Self-similarity and heavy-tailedness are of great importance for network capacity planning purposes in which researchers are interested in developing analytical methods for analysing traffic characteristics. Designers of computing and telecommunication systems are increasingly interested in employing heavy-tailed distributions to generate workloads for use in simulation - although simulations employing such workloads may show unusual characteristics. Congested Internet situations, where TCP/IP buffers start to fill, show long-range dependent (LRD) self-similar chaotic behaviour. Such chaotic behaviour has been found to be present in Internet traffic by many researchers. In this context, the 'Hurst exponent', H, is used as a measure of the degree of long-range dependence. Having a reliable estimator can yield a good insight into traffic behaviour and may eventually lead to improved traffic engineering. In this paper, we describe some of the most useful mechanisms for estimating the tail index of Internet traffic, particularly for distributions having the power law observed in different contexts, and also the performance of the estimators for measuring the intensity of LRD traffic in terms of their accuracy and reliability.**

Keywords- Self-similarity, LRD, Tail Index, Hurst parameter

I. INTRODUCTION

In the Internet, heavy-tailed distributions have been observed in the context of traffic characterization. Heavy-tails can play an important role in traffic self-similarity. Heavy-tailed distributions characterise long-memory processes, with strong time-dependence structures that vanish very slowly. It has been observed that the Ethernet traffic is characterized by self-similar properties [1] and WAN traffic also exhibits self-similar properties [2] particularly when it is associated with WWW transfers [3]. The condition of self-similarity is observed when the autocorrelation function (ACF) of a time-series declines as a power-law, leading to positive correlations among widely separated observations [4].

When the sizes of files are transferred from a web-server, the distribution is heavy-tailed to a good degree of accuracy meaning that there are a large number of small files transferred but the number of very large files transferred remains significant. The superpositions of samples from heavy-tailed distributions aggregate to form long-range dependent time series. It is necessary to model the heavy-tail traffic so that networks can be provisioned based on accurate assumptions of the traffic that they carry. A heavy-tail distribution can characterise the Internet traffic more accurately as a number of multiplexed sources (e.g. video, audio, web requests, eamil, chat, game, etc.) exhibit the properties of selfsimilarity and LRD.

Distributions having infinite variances are called heavy-tailed with the weight of their tails determined by the parameter $\alpha < 2$ [5]. The properties of heavy-tailed distributions are qualitatively different to commonly used memoryless distributions such as the exponential, normal or Poisson distributions. The research in [2] concludes that such exponentiality assumptions are misleading when exploring the presence of heavy-tailed distributions. Heavy-tailed distributions are ubiquitous in the Internet. Paxson [6] noted wide variability in path characteristics such as losses, round-trip times and bandwidth and high variability is one of the landmarks of heavy-tailed distributions. It is evident [1, 7] that the characteristic of the service process (provided by the webservers, routers etc.) in Internet-related systems is heavy-tailed which affects the complexity of such systems.

Self-similar and *long-range dependent* (*LRD*) characteristics of internet traffic have attracted the attention of researchers since 1994 [1, 8]. It is particularly important to understand the link between self-similar and long-range dependence of traffic and performance of the networks. Thus, in [9], it was observed that the performance of networks degrades gradually with increasing self-similarity, which results in queuing delay and packet loss. The more self-similar the traffic, the longer the average queue size. The queue length distribution is caused by self-similar traffic. The tail of the queue length distribution tends to be higher when the traffic is self-similar, thus resulting in a higher probability of buffer overflow (packet loss). The performance results in [410 show that the degree of self-similarity in the traffic increases as the cell loss and cell delay increase for a certain output port buffer size. The LRD property of the traffic fluctuations has important implications on the performance, design and dimensioning of the network. Self-similarity in packetised data networks can be caused by the distribution of file sizes and by human interactions such as teleconferences, voice chat, online video and games etc.

A number of methods have been proposed to estimate the Hurst parameter. Some of the most popular include:

aggregated variance time (V/T), Rescaled-range (R/S), Higuchi's method, wavelet-based methods and HEAF(2). The various methods demonstrate variable performance.

The remainder of this paper is organised as follows. Section II describes the methods for estimation of the tail index. Section III explores a robust mechanism for estimating the tail index. Section IV highlights the existing estimators for estimating the intensity of long-range dependence. Section V examines the reliability of the methods for estimating the Hurst exponent. Finally we draw conclusions in section VI.

II. METHODS FOR ESTIMATING THE TAIL INDEX

In this section various methods for estimating tail index are described which are used in telecommunication network traffic. The principle for detecting the heavy tailed traffic is that the tail of the distribution decays much more slowly than exponential [11]. In general the Pareto model is widely used as it follows heavy tail distribution. The cumulative distribution for Pareto is

$$F(x) = P[X \le x] = 1 - \left(\frac{\beta}{x}\right)^{\alpha}$$

where β represents the smallest (positive constant) possible value of the random variable and α the shape parameter indicating the tail index. Suppose we have a random sample $X_1,\ldots\ldots\ldots,X_n$ from a distribution F satisfying

$$\overline{F}(x) = P[X > x] = 1 - F(x) = \left(\frac{\beta}{x}\right)^{\alpha}$$

$$= \beta^{\alpha} x^{-\alpha} \approx x^{-\alpha} L(x);\ x \to \infty, \alpha > 0$$

where L is a slowly varying function satisfying

$$\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1$$

A random variable X follows a heavy tailed distribution [4, 12] if $P[X > x] \sim Cx^{-\alpha}$, as $x \to \infty$, $0 < \alpha < 2$. (2.1)

The complementary cdf (ccdf) $\overline{F}(x) = 1 - F(x) = P[X > x]$. where α represents the tail index ; $0 < \alpha < 2$. The presence of heavy-tailed distributions in observed data can be explored by equation (2.1) as follows:

$$Lim_{X \to \infty} \frac{d \log \overline{F}(x)}{d \log(x)} = -\alpha \quad (2.2)$$

which appears as a straight line on log-log axes with slope $-\alpha$ for large x.

A number of log-log complementary distribution (LLCD) plots have been illustrated in [5] to estimate the tail weight. These are plots of the ccdf on log-log axes. Having been plotted in this way, heavy-tailed distributions have the property that follows equation (2.2). The random variable X has infinite mean when $\alpha \le 1$, finite mean but infinite variance when $1 \le \alpha \le 2$ and finite mean and variance when $2 < \alpha$ [13]. For the traffic rate process X, the autocorrelation function satisfies [14]

$$r(k) \approx c\,k^{2H-2};\ as\ k \to \infty,\ 0.5 < H < 1 \quad (2.3)$$

where the Hurst parameter H measures the degree of long-range dependence in X in terms of tail-index α in (2.1) and H is given by $H = (3 - \alpha)/2$.

A basic statistical calibration problem is to estimate the shape parameter α, which is the negative of the index of regular variation. A popular method to estimate α is called the Hill estimator, developed by B. M. Hill [15]. Suppose $X_1,\ldots\ldots\ldots,X_n$ are random variables (e.g. web file sizes) from a distribution F and $X_1 > X_2 > \ldots\ldots\ldots > X_n$ are the order statistics. The Hill estimator of α is

$$\hat{\alpha} = \left\{ \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_i}{X_{k+1}} \right\}^{-1} \quad (2.4)$$

where k is the number of upper order statistics used in the estimation. The Hill plot can be defined as $\{(k, \hat{\alpha}),\ 1 \le k \le n-1\}$ and the index found from a stable region in the graph.

The Hill estimator is the most favourable technique [16] to detect the heavy tailedness of the traffic when the underlying distribution is close to Pareto. The plot may sometimes exhibit excessive bias when the distribution is far from Pareto. In fact, the Hill estimator is designed for the Pareto distribution. The Hill plot is not always informative and the alternative estimators described in the literature give alternative Hill plots abbreviated as AltHill, SmooHill for smoothing Hill plot [16], qq estimator [16, 17] and De Haan's moment estimator [18]. The dynamic qq – estimator [16] is given by

$$\hat{\alpha}_{k,n}^{-1} = \frac{\frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)\log\left(\frac{X_{(i)}}{X_{(k+1)}}\right) - \frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)H_{k,n}}{\frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)^2 - \left(\frac{1}{k}\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)\right)^2}$$

where $H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}}$ (2.5)

The dynamic qq-plot can be obtained by plotting

$\left\{\left(k, 1/\hat{\alpha}_{k,n}^{-1}\right),\ 1 \le k \le n\right\}$, which is similar to the Hill plot.

The moment estimator is defined as

$$H_{k,n}^{(r)} = \frac{1}{k}\sum_{i=1}^{k}\left(\log \frac{X_{(i)}}{X_{(k+1)}}\right)^r \quad (2.6)$$

where $H_{k,n}^{(1)}$ is the Hill estimator and $X_1 > X_2 > \ldots\ldots\ldots > X_n$ are the order statistics from a random sample size of n. Define $r = 1, 2$ and then

$$\hat{\gamma}_n = H_{k,n}^{(1)} + 1 - \frac{1/2}{1 - \left(H_{k,n}^{(1)}\right)^2 / H_{k,n}^{(2)}} \quad (2.7)$$

Then the moment is estimated by plotting $\{k, \hat{\gamma}_n\}$.

In addition, the modified qq plot [19, 11] can be illustrated, which is obtained from the following equation by choosing and fixing *k*.

$$\left\{\left(\log\left(\frac{X_j}{m}\right);\ -\log\left(\frac{j}{k+1}\right)\right);\quad 1 \le j \le k\right\} \tag{2.8}$$

where *m* represents a higher order statistics of a distribution for the samples $X_1,\ldots\ldots\ldots,X_n$, i.e., $m = X_1 \ge X_2 \ge \cdots \cdots \ge X_k$ the order statistics of a distribution. If the data follow approximately Pareto, the plot will look like a straight line with slope α. A least squares line can be fitted through the points with small deviation while computing the slope.

A graphical procedure is introduced in [21], called the Sum plot which suggests a proper value for *k* by using the well-known Hill estimator. The sum plot is given by

$$S_k = \alpha^{-1}[\,k\log(k+1) - \sum_{i=1}^{k}\log i\,] \tag{2.9}$$

$$\text{where } \hat{\alpha}^{-1} = \frac{k}{k-1}H_{k,n} - \frac{1}{k-1}\log X^{(1)} \tag{2.10}$$

$H_{k,n}$ can be found from equation (5). The graph will look like a straight line when plotting S_k against *k* and then the slope is estimated from the least squares line.

III. Exploring the Robust Mechanism for Estimating Tail-Index

In this research, we have analysed six different traffic traces, each of sample length (N) 10000. The traces used in the analysis are EPA, NASA-Jul95, NASA-Aug95, ClarkNet, Saskatchewan and Calgary, all publicly available in [21].

The tail index α from these traffic traces is estimated by several methods. Figures 1 and 2 provide a graphical representation of EPA traffic. Results from other estimates are presented in Table I. An instability of the graph in some region has been observed for NASA-Jul95, NASA-Aug95, ClarkNet and Calgary traffic when plotting the moment estimate of gamma. Clearly the moment estimator is not so informative for these traffic traces. The Dynamic qq (dyn-qq) plot was also somewhat unstable for NASA-Jul95, NASA-Aug95 and Saskatchewan traffic.

Here, a number of order statistics, k=9000 have been chosen for the Static qq (stat-qq) and Sum plots. In most traffic cases, α was found to be less than 2, i.e., there is an infinite variance observed in the traces, which implies the existence of heavy-tailedness in the data traffic. The Sum plot yields an index greater than 2 (i.e., α >2) for NASA-Jul95, NASA-Aug95 and ClarkNet. In particular, the Hill plot, Static qq plot and LLCD plot are in good agreement as they provide close results to each other as shown in Table I.

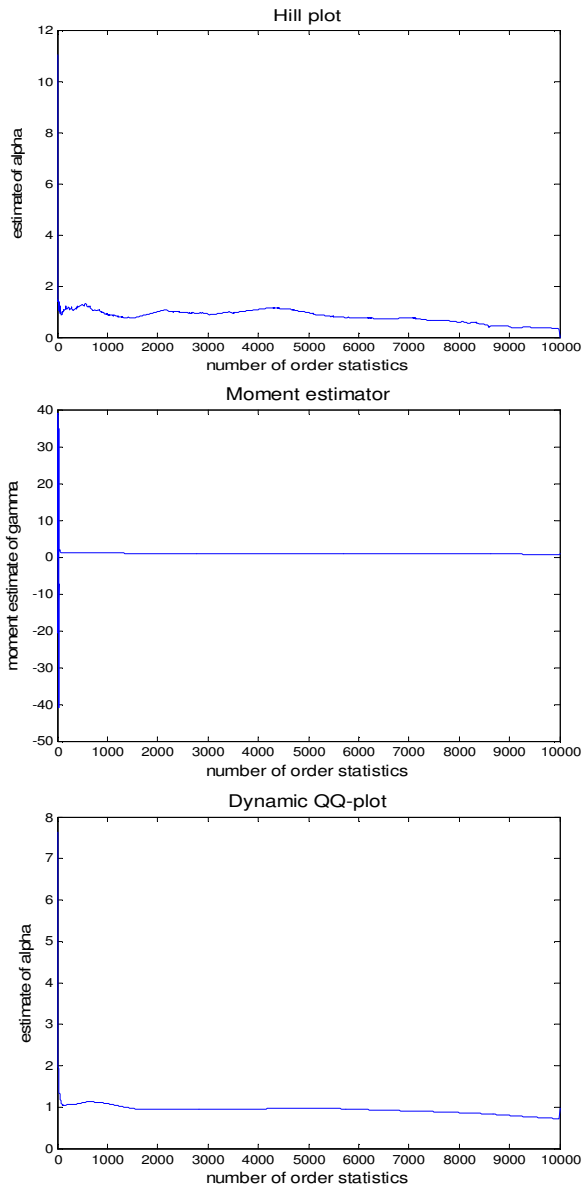

Fig. 1. Estimation of tail index by Hill plot, Dehaan's moment estimator and dynamic-qq plot (EPA-http traffic)

Table I. Estimation of tail index for various http traffic by different methods.

Web File	sample length (N)	Tail index for various methods					
		Hill	moment	dyn-qq	stat-qq	Sum plot	LLCD
EPA	10000	0.764	0.92	0.94	0.74	1.88	0.802
NASA-Jul95	10000	0.583	0.79	1.08	0.57	2.57	0.601
NASA-Aug95	10000	0.619	0.76	0.99	0.60	2.39	0.703
ClarkNet	10000	0.788	1.28	1.11	0.73	2.04	0.810
Saskatchewan	10000	0.830	1.07	1.02	0.82	1.71	0.816
Calgary	10000	0.697	0.80	0.89	0.70	1.76	0.713

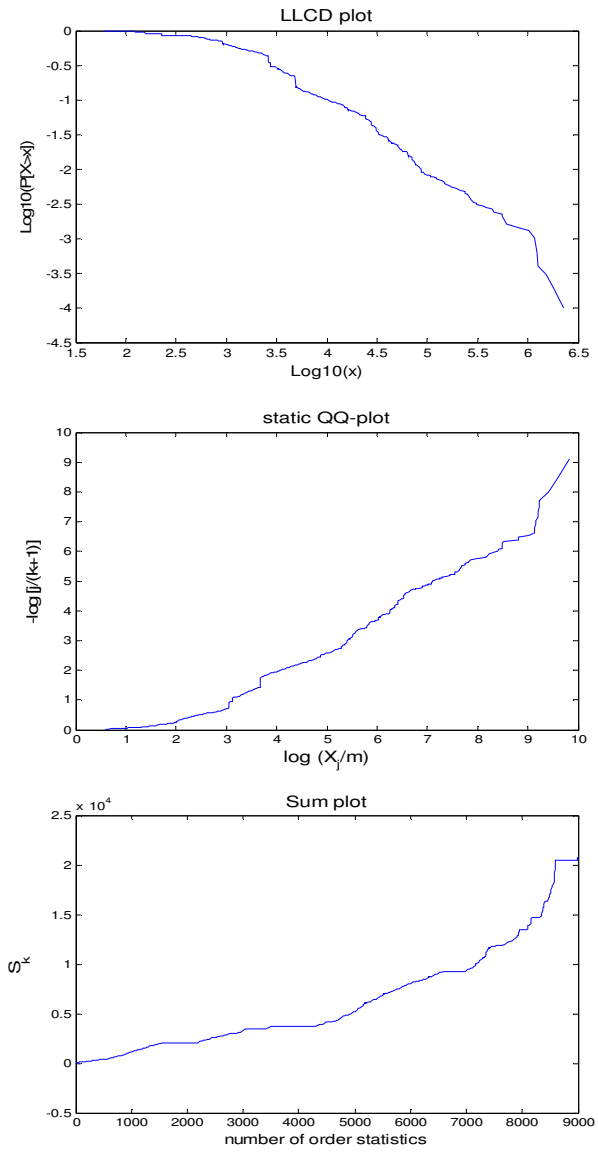

Fig. 2. Estimation of tail index by LLCD plot, static-qq plot and Sum plot (EPA-http traffic)

IV. Estimation of the Hurst Parameter

In this research, we have used five different methods to estimate the Hurst exponent, H:

A. *Variance time (V/T) Analysis*

The self-similarity involves a stationary sequence $X = \{X(i),\ i \geq 1\}$ according to standard time series theory. Let

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i), \quad k = 1,2,3,\ldots\ldots, \quad \left(\frac{N}{m}\right)$$

be the corresponding aggregated sequence with level of aggregation m which can be obtained by dividing the original series X into non-overlapping blocks of size m and averaging over each block. Here the index, k, labels the block. Then the plausible estimator [22] is

$$Var\left(X^{(m)}(k)\right) = \frac{1}{M} \sum_{i=0}^{M-1} \left(X_i^{(m)} - \overline{X}^{(m)}\right)^2$$

where $\overline{X}^{(m)}$ denotes the sample average of $X^{(m)}$ and $\overline{X}^{(m)} = \frac{1}{M} \sum_{i=0}^{M-1} X_i^{(m)}$; Here M is the integer part of N/m.

The estimate for H is found from the estimate of the slope 2H-2 which is measured by fitting a straight line through the points. The Equation (1) is biased in the presence of non-zero correlations, particularly long-range dependence [22]. It is also noted in the literature that if the estimates of the variances are based on few observations (i.e., large m), it becomes unreliable. Another weakness of this method is that the fitting region may arbitrarily be chosen which results in a non-robust estimate.

B. *R/S Analysis*

Self-similar processes are well defined by the rescaled adjusted range (R/S) analysis [1, 23], which is expressed as

$$E\left[\frac{R(n)}{S(n)}\right] \approx c\,n^H, \quad as \quad n \to \infty, \quad where \quad 0.5 < H < 1$$

$$R(n) = \max\left(0, W_1, W_2, \ldots, W_n\right) - \min\left(0, W_1, W_2, \ldots, W_n\right)$$

$$\text{where } W_k = \sum_{i=1}^{k} X_i - k\overline{X}_n; \quad for\ k = 1,2,\ldots, n$$

$$S_n = \text{Samples Standard deviation} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right)}$$

$$\text{and} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The parameter H can therefore be estimated by plotting log (E[R(n)/S(n)]) versus log(n) and measuring the slope by least-square linear approximation technique.

C. *Higuchi method*

The method was proposed by Higuchi [24]. It involves taking the partial sums $Y(n) = \sum_{i=1}^{n} X_i$ (i.e., constructing the cumulative process from the increment process X_i) of the original time series $\{X_i,\ i = 1,2,\ldots\ldots, N\}$. Then, we find the normalized length of the curve, namely

$$L(m) = \frac{N-1}{m^3} \sum_{i=1}^{m} \left[\frac{N-i}{m}\right]^{-1} \sum_{k=1}^{[(N-i)/m]} \left| Y(i+k\,m) - Y(i+(k-1)\,m) \right|$$

where N is the length of the time series, m is essentially a block size and [.] denotes the greatest integer function. Then the estimate of H is found by plotting L (m) versus m in a log-log plot and adding 2 to the slope of the fitted straight line.

D. *Wavelet method*

The wavelet-based estimator was introduced by Abry et al. [25, 26]. The method was based on the generation of the wavelet coefficients and has some attractive properties. So far the method is the most widely used to estimate the long-range dependence parameter in the networking community.

Let ψ_0 denote the mother wavelet. Then construct other wavelets $\psi_{j,k}$ such that

$$\left\{\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - k),\ k \in \mathbf{Z}\right\}$$

Let $d_x(j,k)$ denote the projection of the data set X onto the wavelet $\psi_{j,k}$, namely

$$d_x(j,k) = \left\langle X, \psi_{j,k} \right\rangle;$$

where $\langle .,. \rangle$ denotes the inner product. Then the variance of the wavelet coefficients is estimated by

$$v_j = 2^j \sum_{k=0}^{2^{-j}-1} (d_x(j,k))^2$$

The wavelet estimate of the Hurst parameter is not found by computing v_j for the fractional Brownian motion sample and making a least squares fit of log (v_j) on j. Instead, the following modified method has been applied by performing the same procedure on the wavelet coefficients of the corresponding fractional Gaussian noise sample.

$$\in_j = \log_2(v_j) - (2H - 1)\, j - 2\log_2(\sigma)$$

The slope is then obtained from a linear regression of $\left\{\log_2(v_j) - \mathbf{E}(\in_j)\right\}_{j_1 \le j \le j_2}$ on $\left\{j\right\}_{j_1 \le j \le j_2}$, weighted by $\left\{1/Var(\in_j)\right\}_{j_1 \le j \le j_2}$. The plausible estimator is found by the relationship, slope = 2H-1. The method is implemented in "C" and Matlab programming languages provided by Darryl Veitch [27].

E. HEAF: A 'Hurst Exponent by Autocorrelation Function' Estimator

A new estimator, HEAF, is introduced in [28]. For given observed data X_i (i.e. $X_1, ..., X_n$), the sample autocorrelation function can be calculated by the following method:

$$\text{Let } \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (4.1)$$

$$\text{and } \hat{\gamma}_n(k) = \frac{1}{n}\sum_{i=1}^{n-k}\left(X_i - \hat{\mu}_n\right)\left(X_{i+k} - \hat{\mu}_n\right), \qquad (4.2)$$

where k=0,1, 2, ….., n,

$$\text{with } \hat{\sigma}_n^2 = \hat{\gamma}_n(0). \qquad (4.3)$$

Then the sample autocorrelations of lag k are given by

$$\hat{\rho}_k = \frac{\hat{\gamma}_n(k)}{\hat{\sigma}_n^2} \qquad (4.4)$$

(Equations (4.1), (4.2), (4.3) and (4.4) denote the sample mean, the sample covariance, the sample variance and the sample autocorrelation, respectively). A second-order stationary process is said to be exactly second-order self-similar, with Hurst exponent $1/2 < H < 1$, if

$$\rho_k = 0.5\ [\,(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\,] \qquad (4.5)$$

From equation (4.5), Kettani and Gubner suggest a moment estimator of H. They consider the case k =1 and replace ρ_1 by its sample estimate $\hat{\rho}_1$, as defined in equation (4.4). This gives an estimate for H of the form

$$\hat{H} = \frac{1}{2} + \frac{1}{2\log_e 2}\log_e(1 + \hat{\rho}_1) \qquad (4.6)$$

Clearly, this estimate is straightforward to evaluate, requiring no iterative calculations. For more details of the properties of this estimator, see Kettani and Gubner [29].

An alternative estimator of H is proposed based upon equation (4.5), by considering the cases where *k>1.* Note that the sample equivalent of equation (4.5) can be expressed as

$$f(H) = \hat{\rho}_k - 0.5\{(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\} = 0. \qquad (4.7)$$

Thus, for a given observed $\hat{\rho}_k$, *k>1,* a suitable numerical procedure can be used to solve this equation, and find an estimate of *H*. This is denoted as an HEAF(*k*) estimate of *H*.

To solve equation (4.7) for *H,* the well-known Newton-Raphson (N-R) method is used. This requires the derivative of *f(H)*. Here note that $k \neq 1$,

$$f'(H) = -0.5\left\{\begin{array}{l}(2\log(k+1))(k+1)^{2H} \\ -(4\log(k))(k)^{2H} + \\ (2\log(k-1))(k-1)^{2H}\end{array}\right\} \qquad (4.8)$$

Hence, the algorithm to estimate HEAF(k), for any lag *k,* consists of the following steps:

1. Compute the sample autocorrelations for lag k of a given data set by equation (4.4). (X_i can be denoted as the number of bits, bytes, packets or bit rates observed during the i th interval. If X_i is a Gaussian process, it is known as fractional Gaussian noise).
2. Make an initial approximation for *H*, e.g. H_1 = 0.6, then calculate H_2, H_3, $H_{4,.....}$, successively using $H_{r+1} = H_r - f(H_r)/f'(H_r)$, until convergence, to find the estimate $\hat{H}$ for the given lag *k*. An initial consideration is the case where k = 2 in equation (4.2); i.e. HEAF(2) is considered first.

One of the major advantages of the HEAF estimator is speed, as the N-R-method converges very quickly to a root. There is no general convergence criterion for N-R. Its convergence depends on the nature of the function and on the accuracy of the initial approximation. Fortunately, the form of the function (i.e., equation (4.7)) appears to converge quickly (within at most four iterations) for any initial approximation in the range of interest, namely *H* in *(0.2, 1).* If an iteration value, H_r is such that $f'(H_r) \cong 0$, then one can face "division by zero" or a near-zero number. This will give a large magnitude for the next value, H_{r+1} which in turn stops the iteration. This problem can be resolved by increasing the tolerance parameter in the N-R program. All

HEAF(k), for $k = 2, \ldots, 11$, have been considered and no difficulty in finding the root in (0.5, 1) has been encountered.

V. Exploring the Reliability of the Estimators

In this section, we discuss how to determine a reliable estimator based on simulation experiments. In most cases, researchers use the biased mean square error (mse) and confidence interval (CI) of the estimator to explore its reliability and robustness. Sometimes, however, it is hard to make a decision merely by looking at the CI of the estimator. For instance, for H = 0.7, 100 different realisations of self-similar sequences have been generated, each with sample length N = 10000. For a particular estimator, the CI is found to be (0.583, 0.605). Looking at such a CI for this estimator, one could easily conclude that the estimator outperforms for that particular Hurst parameter. But the real scenario can be observed when looking at the 20 lowest (say) and the 20 highest (say) values of the Hurst parameter out of those 100 realisations. Here, we show a comparison of the reliability of the estimators, such as rescaled-range analysis (R/S), variance-time analysis (V/T), the wavelet-based estimator and Higuchi method in conjunction with the HEAF(2) estimator.

Figure 3 illustrates the performance of the estimators showing the values of Hurst parameters (Hs) for 100 different realisations. For the simulation experiment, we generated the exact self-similar sequences by a fractional Gaussian noise process with the Deitrich-Newsam algorithm [30]. Note that each realisation implies a set of data that contains the sample length, N = 16384. In Figure 3, for H = 0.6, H = 0.7, H = 0.8 and H = 0.9, the wavelet based estimator and HEAF(2) estimator are more stable than the other estimators. However the wavelet based estimator is more stable than HEAF(2) for H = 0.6, H = 0.7 and H = 0.9 when considering the first 50 realisations. For the last 50 realisations, HEAF(2) outperforms the other estimators for H = 0.6, H = 0.7, H = 0.8 and H = 0.9. Note that R/S analysis, V/T analysis and the Higuchi method show a greater fluctuation (i.e. overestimation and underestimation) of Hs for corresponding Hurst parameters.

Figure 4 depicts the performance of the estimators by simulation experiments, with self-similar sequences generated by a FARIMA (0, d, 0) process [31] for a particular Hurst parameter (H). It is clear that the HEAF(2) and wavelet-based methods are more stable than other estimators and that the values of H are in an entirely acceptable range. Again, after the first 3 realisations of Figure 4, HEAF(2) outperforms the other estimators for the corresponding H = 0.6, H = 0.7, H = 0.8 and H = 0.9. Based on the comparison of simulation experiments, for both the fGn and FARIMA (0, d, 0) processes, it is evident that HEAF(2) is a stable method which quantifies the reliable degree of long-range dependence.

It is also evident from the simulation experiment that R/S analysis, V/T analysis and the Higuchi method are not reliable when estimating the LRD traffic, as they sometimes underestimate or overestimate the Hurst values for the corresponding generated realisation (data set) for a particular Hurst parameter.

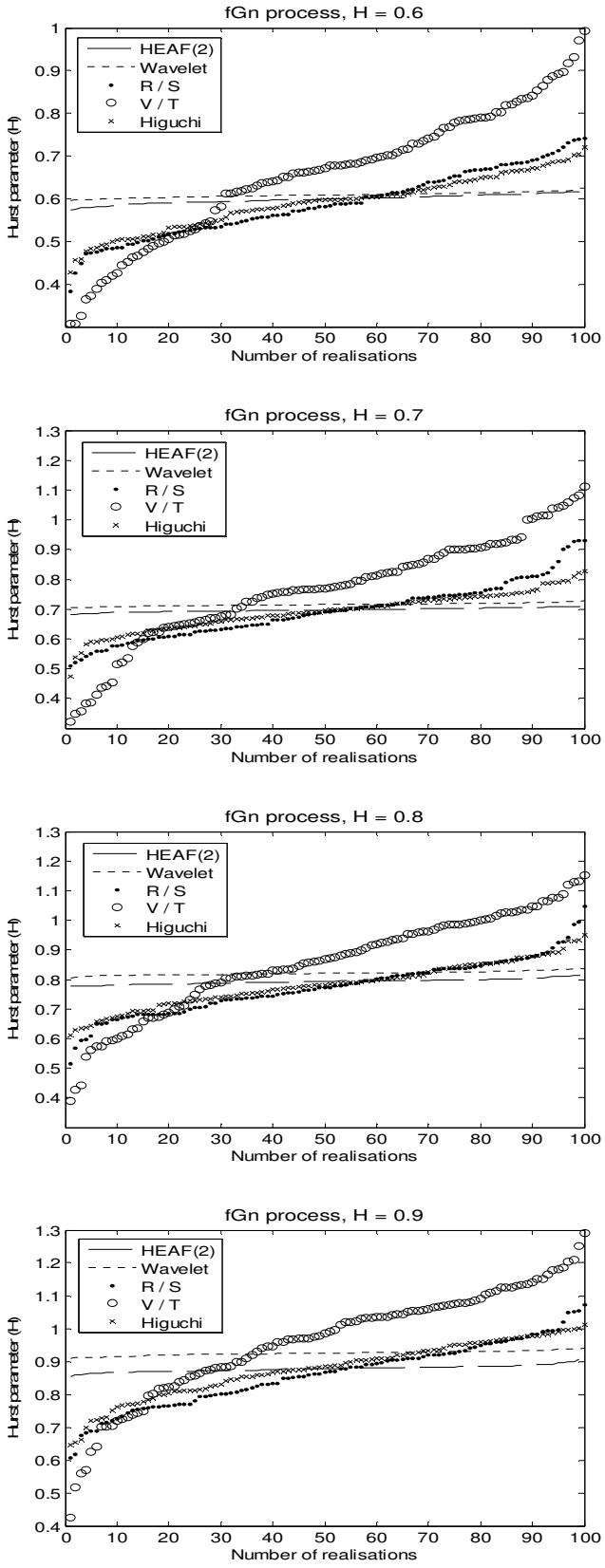

Fig. 3. Reliability of the estimators. A simulation experiment using the fGN process. Sample length N = 16384.

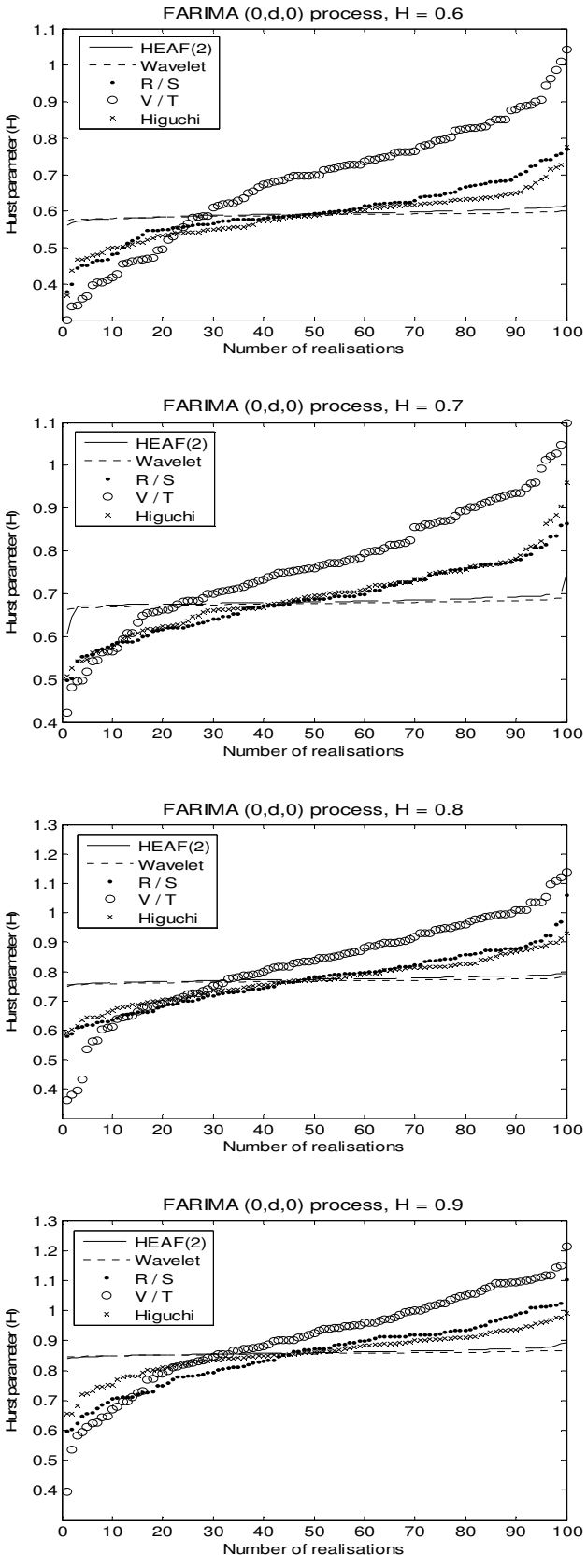

Fig. 4. Reliability of the estimators. A simulation experiment using the FARIMA(0, d, 0) process. Sample length N = 16384.

Table II. 10 lowest Hurst values (out of 100 realisations), fGn process, H = 0.8.

No.	HEAF(2)	Wavelet	R/S	V/T	Higuchi
1	0.777	0.806	0.515	0.388	0.612
2	0.778	0.809	0.568	0.428	0.630
3	0.778	0.809	0.594	0.441	0.635
4	0.779	0.810	0.596	0.537	0.639
5	0.779	0.811	0.609	0.563	0.644
6	0.779	0.812	0.648	0.573	0.658
7	0.780	0.812	0.650	0.573	0.661
8	0.780	0.813	0.651	0.590	0.668
9	0.781	0.813	0.663	0.594	0.671
10	0.781	0.813	0.665	0.599	0.677

Table III. 10 highest Hurst values (out of 100 realisations), fGn process, H = 0.8.

No.	HEAF(2)	Wavelet	R/S	V/T	Higuchi
1	0.805	0.830	0.879	1.048	0.880
2	0.806	0.831	0.885	1.065	0.885
3	0.808	0.831	0.891	1.066	0.886
4	0.809	0.832	0.908	1.076	0.892
5	0.809	0.833	0.924	1.077	0.892
6	0.810	0.833	0.928	1.087	0.904
7	0.811	0.835	0.942	1.119	0.929
8	0.812	0.835	0.986	1.128	0.930
9	0.813	0.835	0.993	1.131	0.932
10	0.814	0.836	1.047	1.152	0.951

Table IV. 10 highest Hurst values (out of 100 realisations), fGn process, H = 0.9.

No.	HEAF(2)	Wavelet	R/S	V/T	Higuchi
1	0.891	0.933	0.984	1.151	0.978
2	0.892	0.934	0.990	1.152	0.982
3	0.892	0.934	0.993	1.164	0.982
4	0.892	0.935	0.994	1.178	0.989
5	0.896	0.937	0.995	1.180	0.990
6	0.897	0.937	1.019	1.186	0.995
7	0.898	0.938	1.050	1.203	0.997
8	0.900	0.939	1.052	1.209	0.999
9	0.906	0.939	1.054	1.251	1.001
10	0.909	0.941	1.073	1.291	1.011

Table V. 10 lowest Hurst values (out of 100 realisations), FARIMA (0,d,0) process, H = 0.8.

No.	HEAF(2)	Wavelet	R/S	V/T	Higuchi
1	0.755	0.749	0.579	0.362	0.591
2	0.755	0.754	0.589	0.380	0.604
3	0.758	0.755	0.609	0.394	0.613
4	0.758	0.755	0.611	0.432	0.636
5	0.759	0.756	0.617	0.536	0.644
6	0.761	0.757	0.617	0.563	0.645
7	0.761	0.757	0.626	0.565	0.645
8	0.762	0.758	0.628	0.602	0.650
9	0.762	0.758	0.628	0.608	0.662
10	0.762	0.758	0.634	0.610	0.666

Table VI. 10 highest Hurst values (out of 100 realisations), FARIMA (0,d,0) process, H = 0.9.

No.	HEAF(2)	Wavelet	R/S	V/T	Higuchi
1	0.873	0.864	0.991	1.098	0.938
2	0.874	0.864	0.995	1.101	0.947
3	0.875	0.865	1.008	1.104	0.948
4	0.875	0.865	1.009	1.109	0.955
5	0.876	0.865	1.012	1.110	0.958
6	0.877	0.865	1.012	1.116	0.963
7	0.878	0.865	1.014	1.116	0.968
8	0.883	0.866	1.018	1.144	0.976
9	0.889	0.866	1.025	1.149	0.978
10	0.892	0.867	1.104	1.215	0.992

The performance of estimators is outlined in Tables II to VI. These tables essentially reflect Figures 3 and 4. However the tables (from II to VI) provide a clearer view of the performance of the estimators in terms of their reliability. For reasons of space limitation we only provide the simulation results for the 10 lowest and 10 highest values of H, while simulating 100 realisations (dataset) for H = 0.8 and H = 0.9. Clearly, in most cases, HEAF(2) outperforms the other estimators.

VI. CONCLUSIONS

The performance of several estimators of the tail index for heavy-tailed Internet traffic have been studied in this research. In most cases, the moment estimator, dynamic qq plot and sum plot are unable to provide an acceptable measured index due to an unstable region observed in the graph. The Hill plot, static qq plot and LLCD plot show a good level of agreement when estimating the index from graphs. Our results show that there are infinite variances (i.e. $\alpha < 2$) observed in the traffic, which is indicative of the existence of heavy-tailedness in Internet traffic.

It has been found that established estimators for the Hurst parameter (with the exception of the wavelet method) can give poor estimates, as they sometimes underestimate or overestimate the degree of self-similarity. For example, for the simulation of $H = 0.6$ and $H = 0.8$, the estimated H values by R/S analysis were found to be 0.38 and 1.059 respectively. (Due to space limitations we cannot provide an exhaustive report our simulation results here.). This is significant because it is possible to derive wrong conclusions and wrong models when measuring the intensity of the LRD with unreliable estimators. Also, based on the comparison of simulation experiments for both fGn and FARIMA (0, d, 0) processes, it is evident that HEAF(2) is a stable method that quantifies the reliable degree of long-range dependence. Through its simplicity, capability of yielding quick estimation, robustness and reliability, we believe that HEAF(2) can be used to estimate the intensity of LRD in real time network traffic.

REFERENCES

[1] Leland Will E. Taqqu M. S., Willinger W. and Wilson D. V., On the Self-similar nature of Ethernet Traffic (Extended version), *IEEE/ACMTransactions on Networking*, February 1994, Vol. 2, No. 1, pp. 1-15.

[2] V. Paxson and S. Floyd., Wide Area Traffic: The Failure of Poisson Modeling, *IEEE/ACM Transactions on Networking*, June 1995, pp.236-244.

[3] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, December 1997, pp.835-846.

[4] Mark E. Crovella and Lester Lipsky, "Long-Lasting Transient Conditions in Simulations with Heavy-Tailed Workloads", In *Proceedings of the 1997 Winter Simulation Conference*, 1997. pp.1005-1012.

[5] Mark E. Crovella and Azer Bestavros, **"**Explaining World Wide Web Traffic Self-Similarity", October 12, 1995, Boston University, Technical Report TR-95-015.

[6] Vern Paxson, End-to-End Internet Packet Dynamics, *IEEE/ACM Transactions on Networking*, June 1999, Vol.7, No.3, pp. 277-292.

[7] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web site", Technical Report, Hewlett-Packard Laboratories, September 1999.

[8] Thomas Karagiannis, Mart Molle, Michalis Faloutsos, Long-range dependence: Ten years of Internet traffic modeling, *IEEE Internet Computing*, 2004, Vol. 8, No. 5, pp. 57-64.

[9] K. Park , G. Kim and M.E. Crovella, On the Relationship Between File Sizes Transport Protocols and Self-Similar Network Traffic, Int'l Conf. Network Protocols, *IEEE CS Press*, 1996, pp. 171-180.

[10] Chen Y., Z. Deng, and C. Williamson, A Model for Self-Similar Ethernet LAN Traffic: Design, Implementation, and Performance Implications, *In Proceedings of the 1995 Summer Computer Simulation Conference*, July 24-26, 1995, pp. 831-837, Ottawa, Canada.

[11] Trang Dang D., Sandor M. and Vidacs A., "Investigation of Fractal properties in Data traffic", *Journal on communications*, 1999, XLIX: 12-18.

[12] Mark E. Crovella, Murad S. Taqqu and Azer Bestavros Heavy-Tailed Probability Distributions in the World Wide Web, In: Robert J. Adler, Raisa E. Feldman, Murad S. Taqqu (eds.), *A Practical Guide To Heavy Tails*, 1998, 1, pp.3-26. Chapman and Hall, New York.

[13] Judith L. Jerkins and Jonathan L. Wang, "From the Network Measurement Collection to Traffic Performance Modeling: Challenges and Lessons Learned", *Journal of Brazilian Computer Society*, vol. 5, No. 3, 1999.

[14] R. H. Riedi and W. Willinger, "Towards an Improved understanding of Network Traffic Dynamics", Self-similar Network Traffic and Performance Evaluation, Wiley, 2000, Chapter20, Eds. Park and Willinger, pp. 507-530.

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

[16] Resnick S. I. , "Heavy Tail Modeling and Teletraffic data". *The Annals of Statistics*, 1997, vol.25, No.5, pp. 1805-1849.

[17] Kratz M. and Resnick S., "The qq Estimator and Heavy tails". *Stochastic Models*, 1996, 12, pp. 699-724.

[18] Dekkers A., Einmahl J. and De Haan, L., "A Moment Estimator for the Index of an Extreme Value Distribution", *The Annals of Statistics*, 1989, vol. 17, pp. 1833-1855.

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

[20] Bruno C. Sousa, "*A Contribution to the Estimation of the Tail index of Heavy-tailed Distributions*", PhD thesis, The University of Michigan, 2002.

[21] Internet traffic archive: http://ita.ee.lbl.gov/html/traces.html

[22] Ton Dieker, *Simulation of fractional Brownian motion*, Masters Thesis, Department of Mathematical Sciences, University of Twente, The Netherlands, 2004.

[23] Hurst H. E., Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers*, 116, pp 770-808, 1951.

[24] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, *Physica D*, 31pp. 277-283, 1988.

[25] P. Abry, P. Flandrin, M. S. Taqqu and D. Veitch, Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data, *Self-Similar Network Traffic and Performance Evaluation*. K. Park and W. Willinger (editors), John Wiley & Sons, New York, 2000.

[26] P. Abry and D. Veitch , Wavelet Analysis of Long-Range Dependent traffic. *IEEE Transactions on Information Theory*, 44(1):2-15, 1998.

[27] http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html

[28] Karim M. Rezaul , Algirdas Pakštas, Robert Gilchrist, Thomas M. Chen, HEAF: A Novel Estimator for Long-Range Dependent Self-similar Network Traffic, Y. Koucheryavy, J. Harju, and V.B. Iversen (Eds.): Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 29 - June 2, 2006, LNCS 4003, pp. 34 – 45.

[29] H. Kettani and J. A. Gubner, A Novel Approach to the Estimation of the Hurst Parameter in Self-Similar Traffic, *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks* (LCN 2002), Tampa, Florida, November, 2002, pp.160-165.

[30] C.R. Dietrich and G.N. Newsam, Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix, *SIAM Journal on Scientific Computing*, 1997, vol.18, pp.1088-1107.

[31] Kokoszka, P. S. and Taqqu, M. S., Fractional ARIMA with stable innovations, *Stochastic Processes and their Applications*, 1995, vol.60, pp.19–47.