4-1-2007

Robustness of HEAF(2) for Estimating the Intensity of Long-Range Dependent Network Traffic

Karim Mohammed Razaul

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

Follow this and additional works at: http://epubs.glyndwr.ac.uk/cairo

Part of the Computer and Systems Architecture Commons, Digital Communications and Networking Commons, Hardware Systems Commons, and the Systems and Communications Commons

Recommended Citation

This Conference Paper is brought to you for free and open access by the Computer Science at Glyndwr University Research Online. It has been accepted for inclusion in Computing by an authorized administrator of Glyndwr University Research Online. For more information, please contact d.jepson@glyndwr.ac.uk.
Robustness of HEAF(2) for Estimating the Intensity of Long-Range Dependent Network Traffic

Abstract
The intensity of Long-Range Dependence (LRD) for communications network traffic can be measured using the Hurst parameter. LRD characteristics in computer networks, however, present a fundamentally different set of problems in research towards the future of network design. There are various estimators of the Hurst parameter, which differ in the reliability of their results. Getting robust and reliable estimators can help to improve traffic characterization, performance modelling, planning and engineering of real networks. Earlier research [1] introduced an estimator called the Hurst Exponent from the Autocorrelation Function (HEAF) and it was shown why lag 2 in HEAF (i.e. HEAF (2)) is considered when estimating LRD of network traffic. This paper considers the robustness of HEAF(2) when estimating the Hurst parameter of data traffic (e.g. packet sequences) with outliers.

Keywords
ACF, HEAF(2), LRD, Self-similarity

Disciplines
Computer and Systems Architecture | Digital Communications and Networking | Hardware Systems | Systems and Communications

Comments
Copyright © ICST 2007. This is the author’s version of the work. It is posted here by permission of ICST for personal use. Not for redistribution. Copyright © 2007 IEEE. Paper presented to the Second IEEE Information and Communication Technologies International Conference (ICTIS'07), held at Fez, Morocco, 3rd -5th April 2007. This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of Glyndŵr University Wrexham's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

This conference paper is available at Glyndŵr University Research Online: http://epubs.glyndwr.ac.uk/cair/60
Robustness of HEAF(2) for Estimating the Intensity of Long-range Dependent Network Traffic

Karim Mohammed Rezaul, Vic Grout

Abstract—The intensity of Long-Range Dependence (LRD) for communications network traffic can be measured using the Hurst parameter. LRD characteristics in computer networks, however, present a fundamentally different set of problems in research towards the future of network design. There are various estimators of the Hurst parameter, which differ in the reliability of their results. Getting robust and reliable estimators can help to improve traffic characterization, performance modelling, planning and engineering of real networks. Earlier research [1] introduced an estimator called the Hurst Exponent from the Autocorrelation Function (HEAF) and it was shown why lag 2 in HEAF (i.e. HEAF (2)) is considered when estimating LRD of network traffic. This paper considers the robustness of HEAF(2) when estimating the Hurst parameter of data traffic (e.g. packet sequences) with outliers.

Index Terms—ACF, HEAF(2), LRD, Self-similarity.

I. INTRODUCTION

The Long-Range Dependence (LRD) property of traffic fluctuations has important implications on the performance, design and dimensioning of the network [2]. A simple, direct parameter characterizing the degree of long-range dependence is the Hurst parameter. The Hurst exponent (or Hurst parameter, $H$), which more than a half-century ago was proposed for analysis of long-term storage capacity of reservoirs [3], is nowadays used to measure the intensity of LRD in network traffic. A number of methods have been proposed to estimate the Hurst parameter. Some of the most popular include the aggregated variance time (V/T) [4], Rescaled-range (R/S) [2, 3], Higuchi method [5], wavelet-based method [6, 7] although there are many others. In all these methods, $H$ is calculated by taking the slope of a log-log plot. So far the wavelet-based Hurst parameter has acquired popularity in estimating LRD traffic. However the study [8] explores the advantages and limitations of wavelet estimators and found that a traffic trace with a number of deterministic shifts in the mean rate results in steep wavelet spectrum which leads to overestimating the Hurst parameter. The intensity of long-range dependence is measured for file size or document size [9], packet counts (number of packets per unit time) [10, 11, 12], interarrival time [13, 14], frame size [15], connection size [16], packet length [17], number of bytes per unit time [2], Bit or byte rate [18] and so on.

This paper continues work on the new estimator introduced earlier which is named Hurst Exponent by Autocorrelation Function (HEAF) [1]. HEAF estimates $H$ by a process which is simple, quick and reliable. In order to investigate the robustness of HEAF(2), two different types of simulation studies were performed. The first one is using fractional Gaussian noise (fGN) sequences generated by the Dietrich-Newsam algorithm [19, 20], which generates exact self-similar sequences. The second one is using a fractional autoregressive moving average (FARIMA) process [21, 22]. Stationarity is assumed for these kinds of classical models (FARIMA and fGN) because it is convenient from a theoretical point of view, especially to check the validity of any hypothesis. The sequences generated by these models show a bell-shaped (i.e. Gaussian) curve either exactly or with small variation. However, our concern in this research is to determine whether, in the case that the underlying process is not FARIMA or fGN, HEAF(2) can still capture the long-range dependency of the traffic. We investigate what role HEAF(2) can play to yield an estimate with a good degree of accuracy if the traffic is nonstationary. For instance, if the data traffic possesses outliers, we consider how to estimate $H$ by eliminating these outliers to have satisfactory and reliable information.

The paper is organised as follows. Section II describes the definitions of self-similarity, long-range dependence and autocorrelation function. Section III introduces the HEAF estimator. Section IV describes about the robust versions of autocorrelation function. Finally the results are presented in section V.

II. SELF-SIMILARITY, LONG-RANGE DEPENDENCE AND AUTOCORRELATION FUNCTION

In general two or more objects having the same characteristics are called self-similarity. A phenomenon that is self-similar looks the same or behaves the same when viewed at different degrees of magnification or different scales on a dimension and bursty over all time scales. Self-similarity is the property of a series of data points to retain a pattern or appearance regardless of the level of granularity used and is the result of long-range dependence in the data series. If a self-similar process is bursty at a wide range of timescales, it may exhibit long-range dependence. In general lagged autocorrelations are used in time series analysis for empirical stationary tests. Self-similarity manifests itself as long-range dependence (i.e., long memory) in the time series of arrivals. The evidence of very slow, linear decay in the sample lag autocorrelation function (ACF) indicates the nonstationary behaviour [23]. The research [24] show that Internet traffic is nonstationary.

Long-range-dependence means that all the values at any time are correlated in a positive and non-negligible way with values at all future instants. For a continuous
time process $Y = \{Y(t), t \geq 0\}$ is self-similar if it satisfies the following condition [25]:

$$y(t) = a^{-H} y(at), \quad \forall a > 0, \quad 0 < H < 1$$

where $H$ is the index of self-similarity, called Hurst parameter and the equality is in the sense of finite-dimensional distributions.

The stationary process $X$ is said to be a long-range dependent process if its autocorrelation function (ACF) is non-summable [26] meaning that

$$\sum_{k=-\infty}^{\infty} \rho_k = \infty$$

The details of how ACF decays with $k$ are of interest because the behaviour of the tail of ACF completely determines its summability. According to [3], $X$ is said to exhibit long-range dependence if its autocorrelation function (ACF) is non-summable [26] meaning that

$$\sum_{k=-\infty}^{\infty} \rho_k = \infty$$

Equation (2.1) implies that the LRD is characterized by an autocorrelation function that decays hyperbolically rather than exponentially fast.

III. HEAF: A ‘HURST EXPONENT BY AUTOCORRELATION FUNCTION’ ESTIMATOR

A new estimator has been introduced [1] by extending the approach of Kettani and Gubner [27]. As in [27], for a given observed data $X_i$ (i.e. $X_1, \ldots, X_n$), the sample autocorrelation function can be calculated by the following method:

Let $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ (3.1) and $\hat{\gamma}_n(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} (X_i - \hat{\mu}_n) (X_{i+k} - \hat{\mu}_n)$, (3.2)

where $k=0, 1, 2, \ldots, n$, with $\sigma_n^2 = \hat{\gamma}_n(0)$.

Then the sample autocorrelations of lag $k$ are given by

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

(Equations (3.1), (3.2), (3.3) and (3.4) 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

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

From equation (3.5), Kettani and Gubner suggest a moment estimator of $H$. They consider the case where $k = 1$ and replace $\hat{\rho}_1$ by its sample estimate $\hat{\rho}_1$, as defined in equation (3.4). This gives an estimate for $H$ of the form

$$\hat{H} = \frac{1}{2} \log \frac{1}{\rho_1}$$

Clearly, this estimate is straightforward to evaluate, requiring no iterative calculations. For more details of the properties of this estimator, see Kettani and Gubner [27].

An alternative estimator of $H$ is proposed based upon equation (3.5), by considering the cases where $k > 1$. Note that the sample equivalent of equation (3.5) can be expressed as

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

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 a HEAF($k$) estimate of $H$.

To solve equation (3.7) for $H$ the well-known Newton-Raphson (N-R) method is used. This requires the derivative of $f(H)$. Here note that $k > 1$,

$$f'(H) = -0.5(2\log(k+1)(k+1)^{2H} - (4\log(k)(k)^{2H} + (2\log(k-1)(k-1)^{2H})$$

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 (3.4). (Note that $X_i$ 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 guess of $H$, e.g. $H_1 = 0.6$, then calculate $H_2, H_3, \ldots$, 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 of the case where $k = 2$ in equation (3.2); i.e. HEAF(2) is considered first.

One of the major advantages of the HEAF estimator is speed, as the NR-method converges very quickly to a root. There is no general convergence criterion for NR. Its convergence depends on the nature of the function and on the accuracy of the initial guess. Fortunately the form of the function (i.e., equation (3.7)) appears to converge quickly (within at most four iterations) for any initial guess in the range of interest, namely $H$ in $(0.2, 1)$. If an iteration value, $H_r$, is such that $f'(H_r) = 0$, 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. A HEAF($k$), for $k = 2, \ldots, 11$, have been considered and no difficulty in finding the root in $(0.5, 1)$ have been encountered.

IV. ROBUST AUTOCORRELATION FUNCTION

The forecasting of network traffic and Quality of Service (QoS) can be affected by the additive outliers. The sample ACF that was used in HEAF (2) is sometimes controversial. In this research we test the performance of HEAF (2) by using three robust ACF such as Trimmed ACF [28], variance-ratio of differences and sums which is known as D/S variance estimator [29, 30], weighted
sample autocorrelation function (shorten as WACF) [31]. Polasek [32] showed in his paper how to eliminate these additive outliers by different robust acf. According to his findings the sample acf (i.e. moment based) was surprisingly ranked as 2 after TACF for eliminating additive outliers. Due to space limitation we only present the results from Trimmed ACF.

The Trimmed ACF can be calculated by the following procedure:

Let $z_1 \leq z_2 \leq \ldots \leq z_n$ be the ordered observations of the given time series $z_1, z_2, \ldots, z_n$. Chan and Wei [28] introduced the $\alpha$-trimmed sample autocorrelation function (shorten as TACF) defined by

$$\hat{\rho}_T(k) = \frac{\sum_{i=k+1}^{n} (z_i - \bar{z}(\alpha)) (z_{i-k} - \bar{z}(\alpha)) L_{i-k}^{(\alpha)} L_{i}^{(\alpha)}}{\sum_{i=k+1}^{n} L_{i-k}^{(\alpha)} L_{i}^{(\alpha)}}$$

where

$$\bar{z}(\alpha) = \frac{1}{\alpha} \sum_{i=1}^{\lceil \alpha n \rceil} z_i$$

and

$$L_{i}^{(\alpha)} = \begin{cases} 0, & \text{if } z_i \leq z_{(g)} \text{ or } z_i \geq z_{(n-g+1)} \\ 1, & \text{otherwise} \end{cases}$$

where $g = \lfloor \alpha n \rfloor$ is the integer part of $\alpha n$ and $0 \leq \alpha \leq 0.05$. Chan and Wei showed that TACF is, in general, very successful in removing the adverse effect of outliers on the estimation of ACF. The parameter called, automatic alpha can be estimated by trimmean filter (TMF) [33, 34].

The procedure for estimating alpha by TMF is as follows:
1. Sort the data in ascending order.
2. Calculate the parameter $Q$ according to the equation below

$$Q = \frac{U(20\%) - L(20\%)}{U(50\%) - L(50\%)}$$

- where $U(x\%)$ is the average of the upper $x\%$ of the ordered sample and $L(x\%)$ is the average of the lower $x\%$ of the ordered sample.
- $Q$ is a measure of the departure of the distribution contained in the sample from a normal distribution.
- Trim off each tail of the ordered distribution according to the value of the trimmean parameter alpha.

Note that the alpha parameter given in [33, 34] is modified here for estimating a good degree of accuracy when considering network traffic data. The TMF assumes the distribution to be symmetric, but not necessarily Gaussian. For a pure Gaussian distribution of data, 4 percent of the data is trimmed from each tail of the original sorted distribution. For a given segment of time, a maximum of 5 percent of the data is trimmed off each tail.

V. RESULTS AND DISCUSSION

In [1, 35, 36], the results show that HEAF(2) is an estimator of $H$ with relatively good bias and mse, when estimating fractional Gaussian noise or FARIMA processes. Because of its simplicity and reliability it is believed that HEAF (2) can be used for real time network traffic control. Of course, a real process will be unlikely to be exactly an fGn process or even FARIMA process. Indeed, a real process may suffer from a ‘noise’, discrepant values or other outliers. This section presents the robustness of the proposed estimator, HEAF(2), against departures from ideal assumptions.

In order to test the robustness of HEAF (2) we generate some noisy sequences by mixing with the data sequences generated by FARIMA (0, d, 0) and fGn processes for a particular Hurst parameter ($H$). Obviously $H$ will be changed when making noisy data, meaning that the process (FARIMA or fGN) no longer exists as it holds additive outliers. Figure 1 shows a pictorial view of noisy data to be analysed in order to explore the robustness of the HEAF(2).

![Fig. 1. Pictorial view of noisy samples (i.e. data with additive outliers)](image-url)
In Figure 2, the top left figure shows a histogram of FARIMA (0, d, 0) process with $H = 0.6$. The estimated $H$ by HEAF(2) is 0.585. The top right figure represents a histogram for a noisy samples generated by mixing with FARIMA (0,d,0) process for $H = 0.6$ having the sample length, $N = 16384$. The histogram at the bottom of Figure 2 is plotted after elimination of the outliers shown in the top right figure. The estimated $H$ for noisy samples (top right figure) and samples after elimination are 0.8993 and 0.576 respectively. The outliers from noisy samples are eliminated by automatic alpha and then the alpha value is used in TACF which in turn applied in HEAF(2).

In Figure 3, uniform random numbers are chosen to generate FARIMA (0, d, 0) sequences for various Hurst parameters. Due to uniform random function used in the process, FARIMA (0,d,0) generates only positive sequences, which can imitate real Internet packet
sequences. It is clear from the results presented in this research that additive outliers can be removed by applying robust ACF and after elimination of these outliers from different case study, it is evident that HEAF(2) yields a reliable value of H.

VI. CONCLUSION

It is possible to end up with wrong conclusions and wrong models when measuring the intensity of the LRD with unreliable estimators. In this research we have shown that the plausible H for given data can be overestimated or underestimated due to additive outliers possessing in the data. These outliers can be removed by applying robust ACF in HEAF(2) and in this case HEAF(2) yields a consistent and reliable results. Because of the simplicity, robustness and reliability, we believe that HEAF(2) can be used to estimate the intensity of LRD in real time network traffic.

REFERENCES


[29] Polasek W., Robust and Resistant measures for the autocorrelation function, mimeo, Institute of Statistics and Informatics, University of Vienna, 1982.


Karim Mohammed Rezaul was born in Bangladesh on November, 1972. He is a Member of Royal Institution of Naval Architects (RINA), UK and Institution of Engineers Bangladesh (IEB), Bangladesh. In 1998, he graduated in the field of Naval Architecture and Marine Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. In 2001, he received an MSc in Marine Technology from Norwegian University of Science and Technology, Trondheim, Norway.

In February 2002, Mr. Karim was appointed as visiting lecturer in the department of computing, communications Technology and Mathematics at London Metropolitan University and continued until June 2005. Now he is a visiting lecturer in the department of computing at Central College London. His research interests include Network Traffic Engineering, Statistical analyses of data, Heavy tail distribution of network traffic, Long-range dependent network traffic, Modelling network traffic by Fractal and wavelet method, Traffic control mechanism and Stochastic process and probability distribution.

Vic Grout was awarded the BSc(Hons) degree in Mathematics and Computing from the University of Exeter (UK) in 1984 and the PhD degree in Communication Engineering from Plymouth Polytechnic (UK) in 1988.

He has worked in senior positions in both academia and industry for twenty years and has published and presented over 100 research papers. He is currently a Reader in Computer Science at the University of Wales NEWI, Wrexham in the UK, where he leads the Centre for Applied Internet Research (CAIR). His research interests and those of his research students span several areas of computational mathematics, particularly the application of heuristic principles to large-scale problems in network design and management.

Dr. Grout is a Chartered Engineer, Chartered Scientist, Chartered Mathematician and Chartered IT Professional, a Member of the IMA, IET(IEE), ACM, IEEE, IEEE Computer and Communications Societies and a Fellow of the British Computer Society (BCS). He chairs the biennial international conference series on Internet Technologies and Applications (ITA 05 and ITA 07).