

A Comparison of Methods for Estimating the Tail Index of Heavy-tailed Internet Traffic

Karim Mohammed Rezaul
Centre for Applied Internet Research
(CAIR)
University of Wales, NEWI
Plas Coch Campus, Wrexham, UK
morekba786@yahoo.co.uk

Vic Grout
Centre for Applied Internet Research
(CAIR)
University of Wales, NEWI
Plas Coch Campus, Wrexham, UK
v.grout@newi.ac.uk

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. In this paper, we describe some of the most useful mechanisms for estimating the tail index, particularly for distributions having the power law observed in different contexts in the Internet.

I. INTRODUCTION

In the Internet, heavy-tailed distributions have been observed in the context of traffic characterization. It has been observed that the Ethernet traffic is characterized by the self-similar properties [1] and WAN traffic also exhibits self-similar properties [2] especially when it is associated with WWW transfers [3]. The condition of self-similarity is that the autocorrelation function (ACF) of time-series declines like a power-law, leading to positive correlations among widely separated observations [4].

When the sizes of files 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. Heavy-tail distribution can characterise the Internet traffic more accurately as a number of multiplexed sources (e.g. video, audio, web requests, email, chat, game, etc.) exhibit the properties of selfsimilarity and LRD.

In most cases the tail index (α) is measured by the so-called Hill estimator which has been used widely in the applied finance and Economics [5, 6], insurance [7] and telecommunications [1, 6, 8, 9, 10.] literatures. The distributions having infinite variances are called heavy-tailed and the weight of their tails is determined by the parameter $\alpha < 2$ [9]. The properties of heavy-tailed distributions are qualitatively different to commonly used memoryless distributions such as the exponential, normal or Poisson distributions. The research [2] concludes that such exponentiality assumptions mislead to explore the presence

of heavy-tailed distributions. The heavy-tailed distributions are ubiquitous in the Internet. Paxson [11] observed that 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. The distribution of burst sizes, for both ftp and HTTP transfers appears to be heavy-tailed [12] and there is little evidence that interarrival times and transfer times are long-tailed. It is evident [1, 13] that the characteristic of the service process (provided by the Web servers, routers etc.) in Internet-related systems is heavy-tailed which affects the complexity of such systems.

The paper is organised as follows. Section II defines the self-similarity and long-range dependence. Section III describes the methods used for estimating tail index in Internet traffic. Finally the results are presented in section IV.

II. SELF-SIMILARITY AND LONG-RANGE DEPENDENCE

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. Several studies [1, 2, 3, 4, 14] have shown that network traffic often shows self-similarity meaning that network traffic shows remarkable bursts at a wide range of time scales. One of the observations [1] showed that the Ethernet LAN traffic is statistically self-similar and Hurst parameter, H is used to measure the burstiness of the traffic. The traffic burstiness (i.e. large variation of traffic bit rate) occurs due to heavy-tailed traffic. If a self-similar process is bursty at a wide range of time scales, it may exhibit long-range dependence (LRD) and the parameter H lies between 0.5 and 1. Note that LRD does not imply self-similar but the process is both LRD and self-similar for $0.5 < H < 1$.

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 [14]: $Y(t) \stackrel{d}{=} a^{-H} Y(at)$, $\forall a > 0$, and $0 < H < 1$ 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 LRD process if its ACF (r_k) is non-summable [15] meaning that $\sum_{k=-\infty}^{\infty} r_k = \infty$

III. METHODS FOR ESTIMATING 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 [16]. In general the Pareto model is widely used as it follows heavy tail distribution. The cumulative distribution for Pareto is

$$F(x) = P[X \leq x] = 1 - \left(\frac{b}{x}\right)^a$$

where b represents the smallest (positive constant) possible value of the random variable and a the shape parameter indicating tail index. Suppose for a random sample X_1, \dots, X_n from a distribution F satisfying

$$\begin{aligned} \bar{F}(x) &= P[X > x] = 1 - F(x) = \left(\frac{b}{x}\right)^a \\ &= b^a x^{-a} \approx x^{-a} L(x); \quad x \rightarrow \infty, a > 0 \end{aligned}$$

where L is a slowly varying function satisfying

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

A random variable X follows a heavy tailed distribution [4, 6] if $P[X > x] \sim Cx^{-a}$, as $x \rightarrow \infty$, $0 < a < 2$. (1)

The complementary cdf (ccdf) $\bar{F}(x) = 1 - F(x) = P[X > x]$.

where a represents the tail index; $0 < a < 2$. The presence of heavy-tailed distributions in observed data can be explored by equation (1) as follows:

$$\lim_{x \rightarrow \infty} \frac{d \log \bar{F}(x)}{d \log(x)} = -a \quad (2)$$

which appears to be a straight line on log-log axes with slope $-a$ for large x .

A number of log-log complementary distribution (LLCD) plots have been illustrated in [9] to estimate the tail weight. These are plots of the ccdf on log-log axes. Having plotted in this way, heavy-tailed distributions have the property that follows the equation (2). The random variable X has infinite mean when $a \leq 1$, finite mean but infinite variance when $1 \leq a \leq 2$ and finite mean and variance when $2 < a$ [17]. For the traffic rate process X , the autocorrelation function satisfies [18]

$$r(k) \approx ck^{2H-2}; \quad \text{as } k \rightarrow \infty, 0.5 < H < 1 \quad (3)$$

where the Hurst parameter H measures the degree of long-range dependence in X in terms of tail-index a in (1) and H is given by $H = (3-a)/2$.

A basic statistical calibration problem is to estimate the shape parameter a which is the negative of the index of regular variation. A popular method to estimate a is called the Hill estimator, developed by B. M. Hill [19]. Suppose X_1, \dots, X_n are random variables (e.g. web file size) from a distribution F and $X_1 > X_2 > \dots > X_n$ be the order statistics. The Hill estimator of a is

$$\hat{a} = \left\{ \frac{1}{k} \sum_{i=1}^k \log \frac{X_i}{X_{k+1}} \right\}^{-1} \quad (4)$$

where k is the number of upper order statistics used in the estimation. Hill plot can be defined as $\{(k, \hat{a}), 1 \leq k \leq n-1\}$

and then the index can be found from a stable region in the graph.

The Hill estimator is the most favourable technique [10] to detect the heavy tailedness of the traffic when the underlying distribution is close to Pareto. The plot sometimes might exhibit excessive bias while the distribution is far from Pareto. In fact, the Hill estimator is designed for the Pareto distribution. Hill plot is not always informative and the alternative estimators described in the literatures are alternative Hill plot abbreviated as AltHill, SmooHill for smoothing Hill plot [10], qq estimator [10, 20] and De Haan's moment estimator [21]. The dynamic qq - estimator [10] is given by

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

$$\text{where } H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log \frac{X_{(i)}}{X_{(k+1)}} \quad (5)$$

The dynamic qq-plot can be obtained by plotting $\left\{ \left(k, 1/\hat{a}_{k,n}^{-1} \right), 1 \leq k \leq n \right\}$ which is similar to the Hill plot.

The moment estimator is defined as

$$H_{k,n}^{(r)} = \frac{1}{k} \sum_{i=1}^k \left(\log \frac{X_{(i)}}{X_{(k+1)}} \right)^r \quad (6)$$

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

$$\hat{g}_n = H_{k,n}^{(1)} + 1 - \frac{1/2}{1 - (H_{k,n}^{(1)})^2 / H_{k,n}^{(2)}} \quad (7)$$

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

In addition the modified qq plot [7, 16] can be illustrated which is obtained from the following equation by choosing and fixing k .

$$\left\{ \left(\log \left(\frac{X_j}{m} \right); -\log \left(\frac{j}{k+1} \right) \right); \quad 1 \leq j \leq k \right\} \quad (8)$$

where m represents a higher order statistics of a distribution for the samples X_1, \dots, X_n , i.e., $m = X_1 \geq X_2 \geq \dots \geq X_k$ be the order statistics of a

distribution. If the data follow approximately Pareto, the plot will look like a straight line with slope a . A least squares line can be fitted through the points with small deviation while computing the slope.

A graphical procedure is introduced in [22], called the Sum plot which suggests a proper value for k by using the well-known Hill estimator. The sum plot is given by

$$S_k = a^{-1} \left[k \log(k+1) - \sum_{i=1}^k \log i \right] \quad (9)$$

$$\text{where } \hat{a}^{-1} = \frac{k}{k-1} H_{k,n} - \frac{1}{k-1} \log X^{(1)} \quad (10)$$

$H_{k,n}$ 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.

IV. RESULTS AND DISCUSSION

In this research, we have analysed six different traffic traces, each of sample length 10000. The traces used in the analysis are EPA, NASA-Jul95, NASA-Aug95, ClarkNet, Saskatchewan and Calgary, all publicly available in [23].

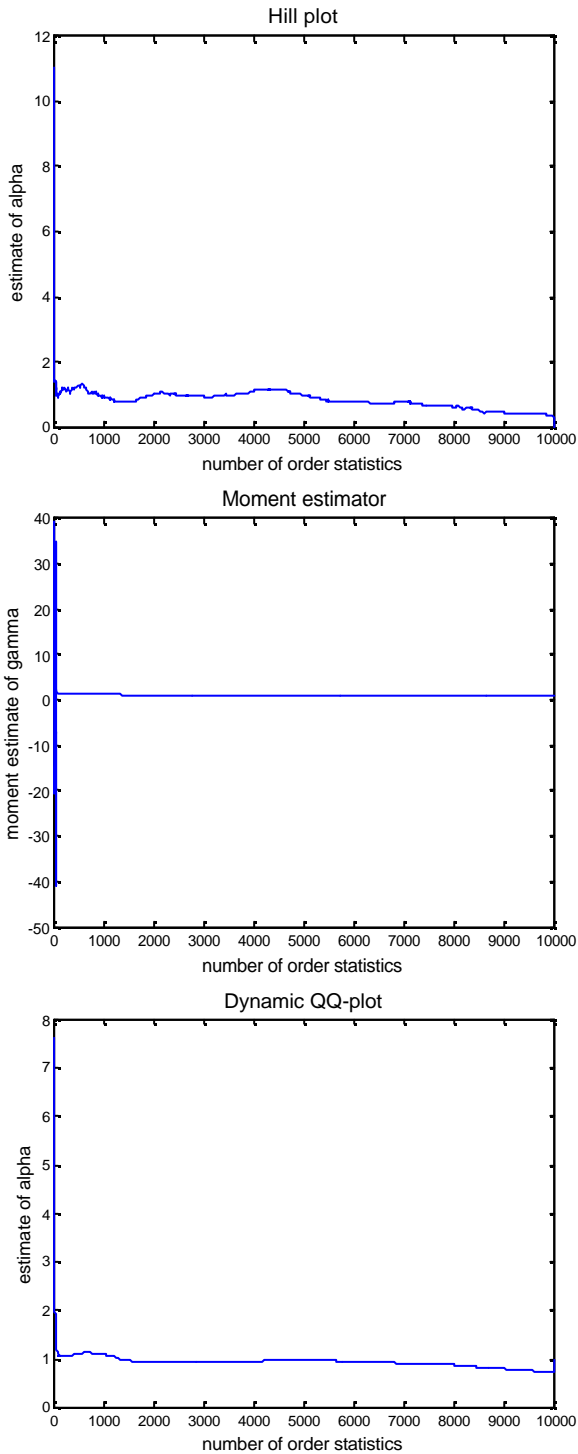


Fig. 1. Estimation of tail index by Hill plot, Dehaan’s moment estimator and dynamic-qq plot (EPA-http traffic)

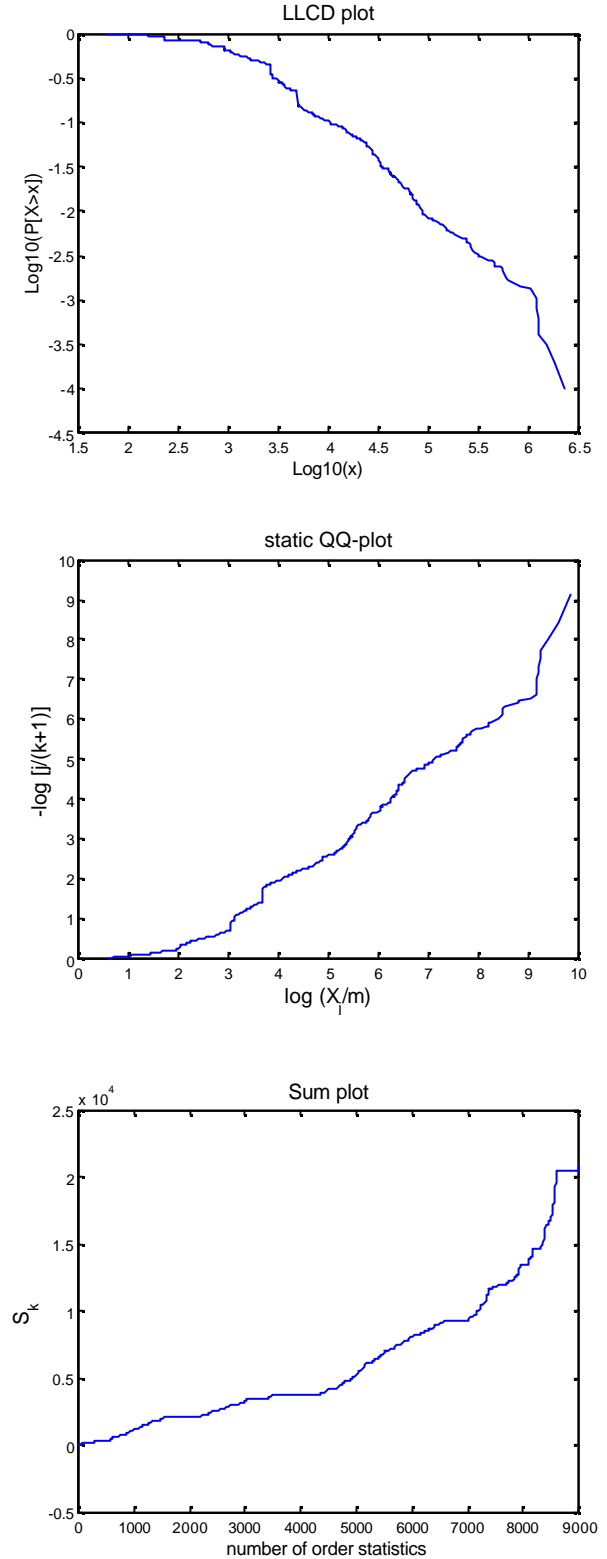


Fig. 2. Estimation of tail index by LLCD plot, static-qq plot and Sum plot (EPA-http traffic)

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 a bit 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., $\alpha > 2$) for NASA-Jul95, NASA-Aug95 and ClarkNet. In particular 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.

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

Web File	I.	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

V. CONCLUSION

The performance of several estimators of tail index for heavy-tailed nature of Internet traffic has 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 because of unstable region observed in the graph. Hill plot, static qq plot and LLCD plot make a good 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 an indicative of the existence of heavy-tailedness in Internet traffic.

REFERENCES

- [1] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, On the Self-Similar Nature of Ethernet Traffic (Extended Version), *IEEE/ACM Transactions on Networking*, February 1994, 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] Jansen, D. and de Vries, C., "On the frequency of large stock returns: putting booms and busts into perspective", *Review of Economics and Statistics*, 1991, 73, pp. 18-24.
- [6] 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.
- [7] P. Embrechts, C. Kluppelberg and T. Mikosh, *Modeling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin Heidelberg, 1997.
- [8] Dacorogna, M. M., M'ller, U. A. and Pictet, O. V., "Heavy-tails in high-frequency financial data", In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, 1998, pp.55-77, Birkhauser
- [9] Mark E. Crovella and Azer Bestavros, "Explaining World Wide Web Traffic Self-Similarity", October 12, 1995, Boston University, Technical Report TR-95-015.
- [10] Resnick S. I. , "Heavy Tail Modeling and Teletraffic data". *The Annals of Statistics*, 1997, vol.25, No.5, pp. 1805-1849.
- [11] Vern Paxson, End-to-End Internet Packet Dynamics, *IEEE/ACM Transactions on Networking*, June 1999, Vol.7, No.3, pp. 277-292.
- [12] Allen B. Downey, "Evidence for Long-tailed Distributions in the Internet", *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, San Francisco, California, USA, 2001, pp. 229 – 241.
- [13] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web site", Technical Report, Hewlett-Packard Laboratories, September 1999.
- [14] Walter Willinger, Vern Paxson, and Murad Taqqu, "Self-similarity and Heavy Tails: Structural Modeling of Network Traffic", In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Adler, R., Feldman, R., and Taqqu, M.S., (editors), , Birkhauser, 1998.
- [15] Cox D., Long-Range Dependence: a Review. H. A. David and H. T. David (eds.), In *Statistics: An Appraisal*, Iowa State Statistical Library, The Iowa State University Press, 1984, pp.55-74.
- [16] Trang Dang D., Sandor M. and Vidacs A., "Investigation of Fractal properties in Data traffic", *Journal on communications*, 1999, XLIX: 12-18.
- [17] 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.
- [18] 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.
- [19] Hill B.M., "A simple approach to inference about the tail of a distribution", *The Annals of Statistics*, 1975, vol. 3, 1163-1174.
- [20] Kratz M. and Resnick S., "The qq Estimator and Heavy tails". *Stochastic Models*, 1996, 12, pp. 699-724.
- [21] 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.
- [22] Bruno C. Sousa, "A Contribution to the Estimation of the Tail index of Heavy-tailed Distributions", PhD thesis, The University of Michigan, 2002.