Log-normal durations can give long range dependence

Jan Hannig
Department of Statistics
Colorado State University
Fort Collins, CO 80523-1877

J. S. Marron
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, New York 14853
and Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Gennady Samorodnitsky
School of Operations Research and Industrial Engineering
and Department of Statistical Science
Cornell University
Ithaca, New York 14853

F. D. Smith
Department of Computer Science
University of North CarolinaChapel Hill, NC
27599-3175

December 21, 2002

Abstract

Duration distributions for internet connections are fit using a novel visualization. While no standard distribution is exactly right, both heavy tail Pareto and light tail log-normal distributions appear sensible in the tails. As noted by Downey (2000), goodness of fit of the log-normal raises interesting questions about the widely accepted view of internet traffic, that only heavy tailed duration distributions lead to long range dependence. Some nonstandard mathematical analysis reveals that both tail distributions are actually consistent with long range dependence, because with appropriate choice of parameters a system with log-normal durations

can have correlation consistent with long range dependence over a wide range of lags.

1 Introduction

A number of studies of internet traffic suggest that internet flows (transfers of data from one computer to another one) often have heavy tailed duration distributions, and that the aggregated traffic (e.g. the collection of all data flowing through a particular point on the internet) exhibits long range dependence, see e.g. Garrett and Willinger (1994) and Paxson and Floyd (1995). An elegant mathematical theory, see e.g. Mandelbrot (1969), Cox (1984), Taqqu and Levy (1986) and Heath, Resnick and Samorodnitsky (1998), provides a convincing connection between these phenomena.

A graphical illustration of this behavior is given in Figure 1, where IP (Internet Packet) flows are represented as horizontal lines. The heights of the lines are random, which allows simple visual separation. Details of the data are given below, but a striking feature is that the lengths of the lines include many very short flows, and also some very long flows.

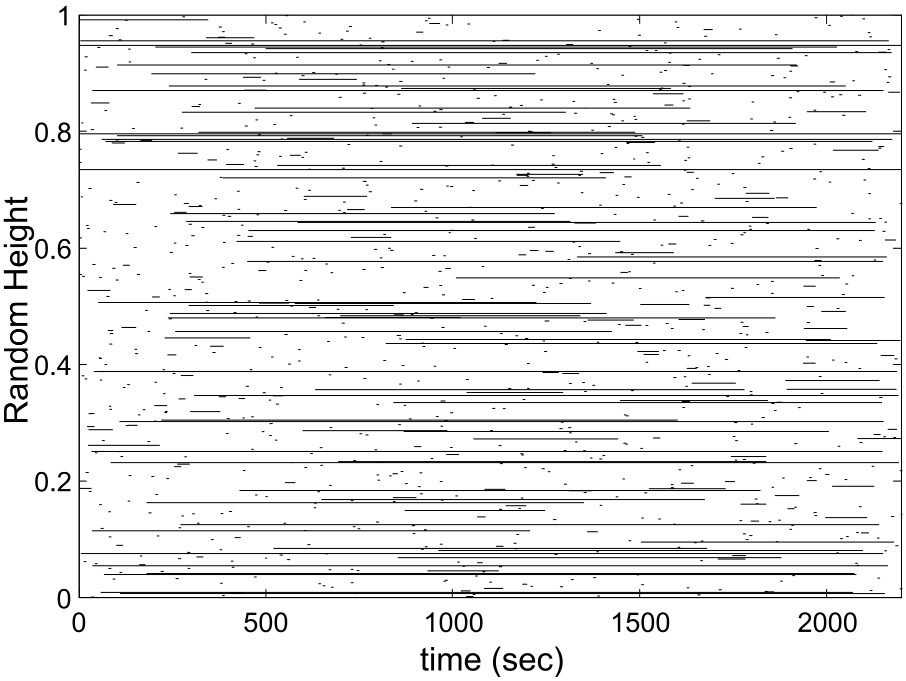

FIGURE 1: *Display of real IP flows, showing "mice" (many short connections) and "elephants" (few long connections), with random vertical "jitter", for convenient visualization.*

The data shown in Figure 1 were gathered from packet headers, during a four hour period on a Sunday morning in 2000, at the main internet link of the University of North Carolina, Chapel Hill. This time period was chosen as being "off peak", having relatively light traffic. An IP "flow" is defined here as the time period between the first and last packets transferred between a given pair of IP sending and receiving addresses. For more details on the data collection and processing methods, see Smith, Hernandez, Jeffay and Ott (2001).

Current popular terminology for the phenomenon of simultaneous occurrence of unusually short and long flows is "mice and elephants". Figure 2 shows a simulation which demonstrates that this duration distribution is far different from the exponential durations that lie at the heart of standard queueing theory. In Figure 2, the flow lengths are randomly drawn from the exponential distribution with the same mean flow length as in Figure 1 (mean = 106 sec.), and the start times are the same as those in Figure 1 . Note that there are far fewer very small flows (mice), and also essentially no very large flows (elephants) in Figure 2, with most of the flows being "medium sized", in stark contrast to Figure 1. This makes it visually clear that the exponential distribution is a very poor approximation to the duration distribution.

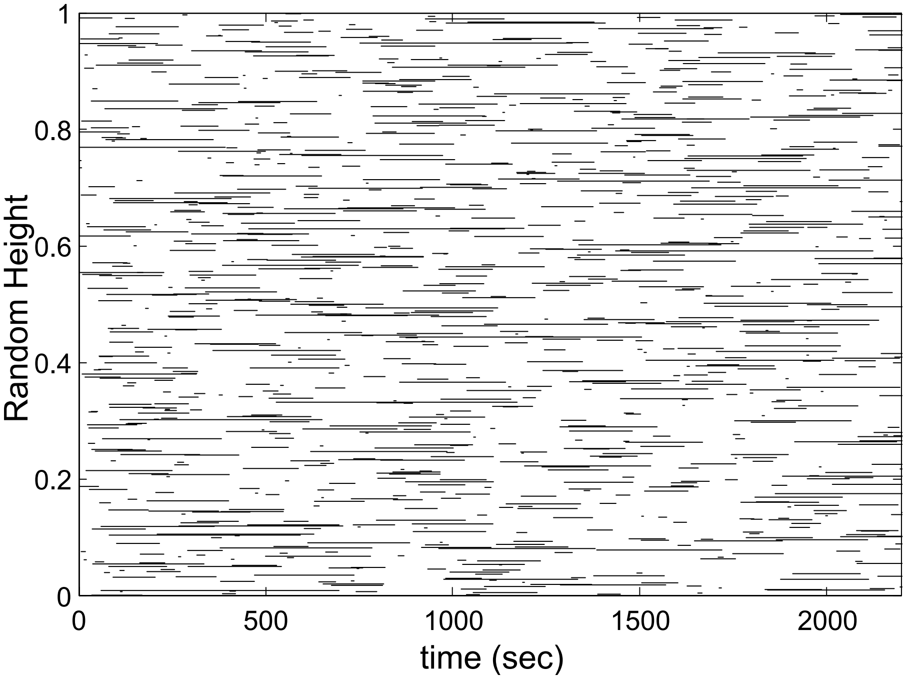

FIGURE 2: *Display of simulated IP flows, with same start times, and same mean duration as in Figure A. Here exponential durations are used, which results in fewer very large, and also fewer very small flows.*

The different duration distributions shown in Figures 1 and 2 lead to far different behavior when the flows are aggregated into a traffic stream, looking either at the sequences of packet time stamps, or else at binned aggregates of either packet counts or packet sizes. In particular, the relatively homogeneous flow lengths in Figure 2 lead to "short range dependent" aggregrations, i.e. autocorrelations which decay exponentially. It may not be surprising that the longer and shorter durations visible in Figure 1 can lead to a different type of dependence structure. In the simplest version of this, see Section 7 of Cox (1984), an infinite variance (i.e. second moment) of the duration distribution implies "long range dependence" in the sense that the autocorrelation decays at a polynomial rate. There are a number of variations on this theme, where the moment condition on the duration distribution is replaced by quantities such as tail indices, and where long range dependence is measured in other ways, such as the behavior of the spectral density near 0. See Chapter 4 of Beran (1994) for discussion of various relations among these.

As noted above, empirical observation of heavy tailed durations and long range dependence, coupled with elegant asymptotic theory connecting them together, suggest that there is a compelling case that we have a deep understanding of internet traffic data. However, Downey (2000) has recently called the depth of this "understanding" into question by some interesting empirical work, which suggests that the log-normal distribution may be more appropriate than classic heavy tailed distributions such as the Pareto. Downey's work is somewhat different from the above, because he analyzes distributions of computer file sizes, which may be somewhat different from IP flows as considered above. However, Downey provides further backing of the log-normal distribution by suggesting an intuitive mechanism which results in log-normal file size distributions. Since the same intuitive mechanism is sensible as well for size distributions of IP flows, the log-normal should be viewed as a serious candidate. At first glance this casts serious doubt on the above empirical plus theoretical view. The reason is that the log-normal has all moments finite, thus failing to have the apparently required infinite second moment. But careful consideration reveals a logical gap: the existing theory uses only one type of asymptotic analysis. It is possible that a finite second moment is consistent with a slow decay of autocorrelations over a particular long range of lags. An important goal of this paper is to fill in this gap by showing that log-normal distributions and long range dependence are consistent with each other, in this sense.

First an empirical study of duration distributions (more relevant than the file size distributions studied by Downey) is given in Section 2. An important feature of the analysis is a novel visualization, which improves previous distributional analyses by providing insight into the level of sampling variability. The analysis shows that neither the Pareto, nor the log-normal provides a particularly good fit. Yet both could be regarded as "acceptable approximations". The Pareto which gives this fit has shape parameter between 1 and 2 (i.e. the first moment is finite, but the second moment is infinite) So which is "right"? Is there a "heavy tailed Pareto" leading to long range dependence, or is there a "light tailed log-normal" leading to short range dependence?

Second a theoretical analysis is given in Section 3, which shows that these two different approximations need not lead to divergent conclusions. In particular, it is seen that there are sequences of log-normal distributions which can yield long range dependence in the sense of polynomially decreasing autocorrelations. The argument is asymptotic in nature, however it should be viewed as saying, in accordance with the above, that a slow decay of correlations is observed over a certain wide range.

2 Fitting of duration distributions

The data set in Figure 1 is not ideal for studying tail behavior of duration distributions because of boundary effects created by the limited time span (about 40 minutes) considered there. In particular, too many flows last for essentially the full time span. Hence, we replace the duration variable by the surrogate variable of flow size. This makes some sense because larger files do require more time to transfer. However, we acknowledge that the approximation is crude. Thus, a set of $n = 734,814$ HTTP response sizes, gathered at the UNC main link in 1998 is considered in this section, entailing that "duration" is now measured in terms of file size, instead of time required for the transfer, as in Figure 1.. Here "flow" also has a somewhat different meaning because these are only the files that are transferred while web browsing, as opposed to all types of data, as considered in Figure 1.

Figure 3 shows a Pareto Q-Q plot analysis of the response size data. This is a graphical technique for assessing the goodness of fit of a probability distribution to data. See e.g. Fisher (1983) for an overview of related techniques. Here the data quantiles (i.e. sorted data values) are plotted as a function of the corresponding theoretical quantiles (the theoretical inverse cumulative distribution function, evaluated at the points $1/(2n), 3/(2n), ..., (2n-1)/(2n)$). If the data come from exactly the theoretical distribution, then the resulting curve (shown as a thick, black solid line) would be close to the 45 degree diagonal line (shown as the black dashed line), except for some random sampling variability.

The region shaded by dotted lines is a visual device for understanding the magnitude of the sampling variability. It is an overlay of Q-Q plots for 100 simulated data sets of size $n = 734,814$ from the theoretical distribution. If the data comes from the theoretical distribution, then with high probability it should lie within the envelope. Large departures from the envelope indicate regions in which the theoretical distribution is a poor fit. In Figure 3 it is apparent that the fit is very poor for small data values. The dotted envelope in Figure 3 is very narrow, especially at the lower end where the 100 curves converge into a single thin line. The reason is that for the relatively large sample size of $n = 734,814$ the natural sampling variability is relatively small.

In Figure 3, both axes are shown on the log scale. This is because for the heavy tailed distributions considered here, only a few large values dominate the whole picture on the ordinary scale. The theoretical distribution shown is the Pareto, with cumulative distribution function

$$F(x) = \left[1 - (x/\sigma)^{\alpha}\right] 1_{(\sigma,\infty)}(x),$$

where the shape parameter $\alpha = 1.24$ and the scale parameter $\sigma = 1499$. The corresponding complementary cumulative distribution function decreases like $x^{-\alpha}$, as $x \to \infty$, so it has an infinite variance, but finite mean. These fitted values were estimated by "quantile matching", in particular they make the empirical and theoretical 0.99 and 0.999 quantiles the same. The location of these two quantiles is shown by two circles in the plot (the thick black curve crosses the 45 degree line at these points). Three other quantiles are indicated by plus signs, to show which parts of the curve represent which parts of the data set.

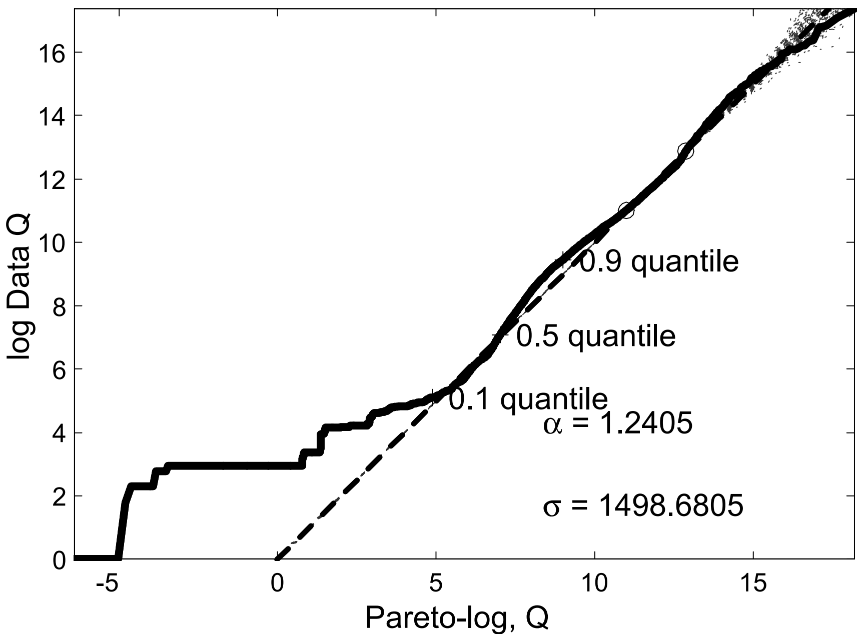

Figure 3: *Pareto Q-Q plot for response size data. Gray envelope gives visual impression of variability. Shows reasonable, but not perfect, fit for larger quantiles.*

Figure 3 shows that the response size data is not perfectly fit by the Pareto(1.24, 1499) distribution. The fit is particularly poor for the smaller data values. But even above the median, the Q-Q curve bends substantially outside of the gray envelope, which shows the distributional shape is also significantly different from the Pareto in that region. However, with such a large data set, it would be surprising if any simple distribution gave a perfect fit. Also no effort had been made to fit the bulk of the distribution, but only the larger values, because these drive the tail behavior being studied here. Furthermore, for the purpose of "heavy tail durations lead to long range dependence", this level of fit appears to be reasonably adequate.

Figure 4 is a parallel analysis, where the underlying theoretical distribution is replaced by the log-normal. Again the parameters of the log-normal, $\mu = 5.28$ and $\sigma = 2.46$ were chosen by 0.99 and 0.999 quantile matching.

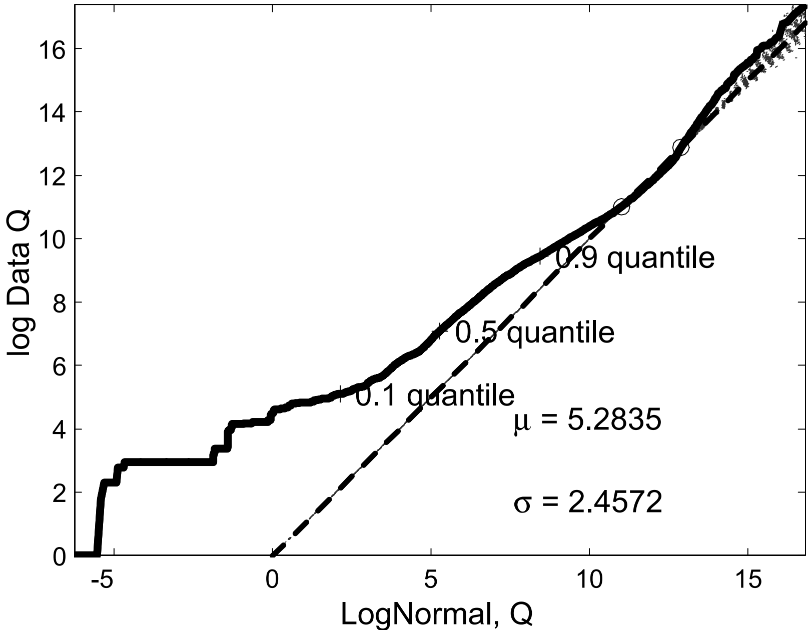

FIGURE 4: *Log-normal Q-Q plot for the response size data. Shows fit is not much worse than for the Pareto.*

Figure 4 shows that similar lessons to those for the Pareto (used in Figure 3) apply. In particular, the fit is not perfect, and is rather poor for small values. But again no effort is made to fit the bulk of the distribution, but instead the emphasis is on the "upper tail", where the fit is reasonably acceptable. There does seems to be "more overall curvature" than for the Pareto, so this distribution does not give quite as good a fit in the bulk of the distribution, but still the fit is surprisingly good in view of the above stated differences between the Pareto(1.24) (heavy tailed, infinite variance), and the log-normal (light tailed, all moments finite). The theoretical results shown in Section 3 show that the light tailed log-normal, can be "heavy tailed enough over a large enough range" to yield the "long range dependence over a broad range of lags" that had previously been associated only with heavy tailed duration distributions.

Many variations are possible concerning the Q-Q analyses done here. To save space, and because the lessons learned are tangential to the main point of this paper, these are not shown here. However some graphics can be found in the web directory

http://www.unc.edu/depts/statistics/postscript/
papers/marron/NetworkData/LogNorm2LRD/

For example, if the 0.99 and 0.999 quantiles are replaced by the 0.9 and 0.999 quantiles, then it is seen in the file RespSize2logNormQQall2.ps that the log-normal yields a substantially better fit in the body of the distribution, at the price of a poorer fit in the upper tail. A wide range of different quantiles for the Pareto can be studied from the movie files RespSize2ParQQq1p5.avi, RespSize2ParQQq1p9.avi, RespSize2ParQQq1p99.avi, RespSize2ParQQq1p999.avi and RespSize2ParQQq1p9999.avi. To see that the Weibull distribution gives a much worse fit than either the Pareto or log-normal considered here, see RespSize2WeibullQQall.ps.

All data analyzed here were kindly provided by the UNC Computer Science Distributed and Real-Time Systems Group, http://www.cs.unc.edu/Research/dirt/.

3 Log-Normal durations give long range dependence

A deliberately simple model for the random process illustrated in Figure 1 is considered here. Many variations are possible, and we view the establishment of similar results in more realistic and general contexts as interesting open problems. For simplicity, only continuous time processes are considered here. A sequence of such models, indexed by $n = 1, 2, ...$ is considered because "heavy tails" and "long range dependence" are asymptotic concepts. The flow arrival process (the point process of starting times of the horizontal line segments in Figure 1) is a standard Poisson process with intensity parameter λ_n. The duration times (the random lengths of the line segments) L_n, are independent, identically distributed, with a log-normal (μ_n, σ_n) distribution independent of the Poisson arrival process. Aggregation of the traffic is represented by $X_{n,t}$, the number of active flows (line segments) at time t.

One way to express long range dependence is in terms of the rate of decay of the autocovariance

$$r(t; \mu_n, \sigma_n, \lambda_n) = \operatorname{cov}(X_{n,s}, X_{n,t+s}).$$

In particular, polynomial decay in t, $r(t) \sim t^{-(\alpha-1)}$ (in the sense that $\lim_{t\to\infty} \frac{r(t)}{t^{-(\alpha-1)}} \in (0,\infty)$) with exponent $\alpha - 1 \in (0, 1)$, is typically viewed as a symptom of long range dependence. This decay is easily obtained if L_n are Pareto, or asymptotically Pareto, because for the above model, the autocovariance is simply and directly related to the tail of the duration distribution, as

$$r(t; \mu_n, \sigma_n, \lambda_n) = \lambda_n \int_t^\infty P(L_n \geq s)\, ds, \tag{1}$$

as seen for example in Cox (1984) and Resnick and Samorodnitsky (1999).

The main goal of this section is to find sequences of parameters μ_n, σ_n, λ_n for which the sequence of processes $X_{n,t}$ exhibits this behavior in the sense that, for a given $C > 0$, for every sequence T_n such that $\log T_n = o\left(n^{1/2}\right)$,

$$\lim_{n\to\infty} \sup_{1\leq t\leq T_n} \left| \frac{r\left(t; \mu_n, \sigma_n, \lambda_n\right)}{Ct^{-(\alpha-1)}} - 1 \right| = 0. \tag{2}$$

This says that the n-th model, is effectively long range dependent over the long range of lags 1 to T_n.

Because the log-normal is "light tailed" according to most classical definitions (e.g. having all moments finite), the key to (2) is to find parameter sequences over which the L_n duration distribution "looks approximately heavy tailed over a wide enough range". This can be done for the log-normal by assuming:

$$\mu_n = -n, \tag{3}$$

$$\sigma_n = \sqrt{\frac{-\mu_n}{\alpha}} = \sqrt{\frac{n}{\alpha}}. \tag{4}$$

Assumption (3) means that most of the mass of the log-normal will be concentrated near 0, i.e. there will be "many mice" (the short line segments in Figure 1). But assumption (4) ensures a "few elephants" (the long lines in Figure 1). Because of the lightness of the tails of the log-normal distribution, a final assumption is needed, to ensure the existence of enough elephants to create long range dependence. This comes from an assumption of "increasing intensity":

$$\lambda_n = \sqrt{\alpha n} e^{\alpha n/2}. \tag{5}$$

At first glance, assumption (5) might seem very strong, however, in an environment of exponentially increasing internet traffic, it is worth contemplating, and is perhaps not far from realistic. We will show that these assumptions give

$$\lim_{n\to\infty} \sup_{1\le t\le T_n} \left| \frac{r\left(t;\mu_n,\sigma_n,\lambda_n\right)}{(2\pi)^{-1/2}\frac{1}{\alpha-1}t^{-(\alpha-1)}} - 1 \right| = 0, \tag{6}$$

which is (2) for the particular

$$C = \frac{1}{(\alpha-1)(2\pi)^{1/2}}.$$

Rescaling λ_n appropriately will give (2) for a general $C > 0$.

To establish (6), note that the integrand of (1) can be rewritten as

$$\begin{aligned} P(L_n &\ge s) = P\left(\exp\left(\mu_n + \sigma_n Z\right) \ge s\right) = P\left(Z \ge \left(\log s - \mu_n\right)/\sigma_n\right) \\ &= P\left(Z \ge \sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}}\log s\right), \end{aligned}$$

where Z is a standard Gaussian random variable. A useful bound (leading to Mill's ratio) comes from the inequalities, valid for any $t > 0$, and following from integration by parts (see e.g. problem 4.14.1.c of Grimmett and Stirzaker (2001))

$$(2\pi)^{-1/2}\left(t^{-1} - t^{-3}\right)e^{-t^2/2} \le P\left(Z > t\right) \le (2\pi)^{-1/2}t^{-1}e^{-t^2/2}. \tag{7}$$

Using (5), and applying the right hand bound in (7) gives, for every $s \geq 1$,

$$\begin{aligned}
\lambda_n P\left(L_n \geq s\right) \quad &\leq \quad \sqrt{\alpha n} e^{\alpha n/2} \left(2\pi\right)^{-1/2} \frac{1}{\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s} e^{-\frac{1}{2}\left(\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s\right)^2} \\
&\leq \quad \left(2\pi\right)^{-1/2} s^{-\alpha} e^{-\frac{\alpha (\log s)^2}{2n}} \leq \left(2\pi\right)^{-1/2} s^{-\alpha}.
\end{aligned}$$

Hence for all $t \geq 1$,

$$r\left(t; \mu_n, \sigma_n, \lambda_n\right) \leq \left(2\pi\right)^{-1/2} \frac{1}{\alpha - 1} t^{-(\alpha-1)}.$$

Similarly using the left hand bound in (7), let S_n be an increasing sequence, such that $\log S_n = o\left(n^{1/2}\right)$. For every $1 \leq s \leq S_n$

$$\begin{aligned}
\lambda_n P(L_n \geq s) \geq & \sqrt{\alpha n} e^{\alpha n/2} (2\pi)^{-1/2} \exp\left[-\frac{1}{2}\left(\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s\right)^2\right] \\
& \times \left[\frac{1}{\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s} - \frac{1}{\left(\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s\right)^3}\right] \\
= & (2\pi)^{-1/2} s^{-\alpha} e^{-\frac{\alpha (\log s)^2}{2n}} \\
& \times \left[\frac{\sqrt{\alpha n}}{\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s} - \frac{\sqrt{\alpha n}}{\left(\sqrt{\alpha n} + \sqrt{\frac{\alpha}{n}} \log s\right)^3}\right] \\
\geq & (2\pi)^{-1/2} s^{-\alpha} c_n,
\end{aligned}$$

where

$$c_n = e^{-\frac{\alpha (\log S_n)^2}{2n}} \left(\frac{1}{1 + \frac{\log S_n}{n}} - \frac{1}{\alpha n}\right)$$

and $\lim_{n\to\infty} c_n = 1$. Consider an increasing sequence T_n, such that $\log T_n = o\left(n^{1/2}\right)$. Clearly if $S_n = T_n^2 + n$ then $\log S_n = o\left(n^{1/2}\right)$. Hence if $1 \leq t \leq T_n$

$$\begin{aligned}
r(t; \mu_n, \sigma_n, \lambda_n) \geq & \int_t^{t^2+n} (2\pi)^{-1/2} s^{-\alpha} c_n ds \\
= & (2\pi)^{-1/2} c_n \frac{1}{\alpha - 1} \left[t^{-(\alpha-1)} - (t^2+n)^{-(\alpha-1)}\right],
\end{aligned}$$

and so

$$\begin{aligned}
\frac{r(t; \mu_n, \sigma_n, \lambda_n)}{(2\pi)^{-1/2} \frac{1}{\alpha-1} t^{-(\alpha-1)}} \geq & c_n \left[1 - \left(\frac{t^2+n}{t}\right)^{-(\alpha-1)}\right] \\
\geq & c_n \left(1 - 2^{-(\alpha-1)} n^{-(\alpha-1)/2}\right).
\end{aligned}$$

Hence,

$$\lim_{n\to\infty} \sup_{1 \leq t \leq T_n} \left|\frac{r(t; \mu_n, \sigma_n, \lambda_n)}{(2\pi)^{-1/2} \frac{1}{\alpha-1} t^{-(\alpha-1)}} - 1\right| = 0.$$

4 Conclusions

This paper considered the controversy of whether internet flow distributions are heavy tailed or not, with a particular view towards understanding the implications for long range dependence. Some data analysis suggested that both the heavy tail Pareto and the light tail log-normal give reasonable fits, although neither is perfect, and the Pareto is somewhat better. This appears contradictory, because the Pareto that fitted had an infinite variance, while the log-normal had all moments finite. Some new theoretical work revealed that these distributions are not so inconsistent as was previously thought, which is consistent with the above data analysis. In particular it is shown here that even (a sequence of suitable parametrizations of) the light tailed log-normal distribution can lead to long range dependence. A clear lesson is that moments (e.g. finiteness of variance) provide a poor way of understanding the type of distributional properties that are important to internet traffic.

Interesting open problems that follow from this work include a corresponding data analysis for other data sets, and generalizations of the theoretical results. Potential generalizations of the theory include finding other parameter sequences for the log-normal giving long range dependence, and an investigation of which other light tailed parametric families can yield long range dependence. There is also lots of room for improvement of modelling of the duration distribution, including mixture and "piece-wise" models, which could then yield parallel theoretical results.

5 Acknowledgement

The research of J. S. Marron was supported by NSF Grant DMS-9971649, and of Gennady Samorodnitsky by NSF grant DMS-0071073.

References

[1] Beran, J. (1994) *Statistics for long-memory processes*, Chapman & Hall.

[2] Cox(1984) Long-Range Dependence: A Review, in *Statistics: An Appraisal, Proceedings 50th Anniversary Conference*. H. A. David, H. T. David (eds.). The Iowa State University Press, 55-74.

[3] Downey, A. B. (2000) The structural cause of file size distributions, Wellesley College Tech. Report CSD-TR25-2000.

[4] Fisher, N. I. (1983) Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography, *International Statistical Review*, 51, 25-58.

[5] Garrett, M. W. and Willinger, W. (1994). Analysis, Modeling and Generation of Self-Similar Video Traffic, *Proc. of the ACM Sigcom '94*, London, UK, 269-280.

[6] Grimmett, G. R. and Stirzaker, D. R. (2001) *Probability and Random Processes*, Oxford University Press, Oxford.

[7] Heath, D., Resnick, S. and Samorodnitsky , G. (1998) Heavy tails and long range dependence in on/off processes and associated fluid models, *Mathematics of Operations Research*, 23, 145-165.

[8] Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994). On the Self-Similar Nature of Ethernet Traffic (Extended Version), *IEEE/ACM Trans. on Networking*, 2, 1-15.

[9] Mandelbrot, B. B. (1969) Long-run linearity, locally Gaussian processes, H-spectra and infinite variance, *International Economic Review*, 10, 82-113.

[10] Paxson, V. and Floyd, S. (1995) Wide Area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, 3, 226-244.

[11] Resnick, S. and Samorodnitsky, G. (1999) Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues, *Queueing Systems*, 33, 43-71.

[12] Smith, F. D., Hernandez, F., Jeffay, K. and Ott, D. (2001) ”What TCP/IP Protocol Headers Can Tell Us About the Web”, *Proceedings of ACM SIGMETRICS 2001/Performance 2001*, Cambridge MA, June 2001, pp. 245-256.

[13] Taqqu, M. and Levy, J. (1986) Using renewal processes to generate LRD and high variability, in: *Progress in probability and statistics*, E. Eberlein and M. Taqqu eds. Birkhaeuser, Boston, 73-89.