Variable Heavy Tailed Durations in Internet Traffic, Part II: Theoretical Implications

F. Hernández-Campos [*] J. S. Marron [†§] G. Samorodnitsky [§‡] F. D. Smith [*]

Abstract

This paper is part of a larger paper that studies tails of the duration distribution of Internet data flows, and their "heaviness". Data analysis motivates the concepts of moderate, far and extreme tails for understanding the richness of information available in the data. The data analysis also motivates a notion of "variable tail index", which leads to a generalization of existing theory for heavy tail durations leading to long range dependence. In this part, the emphasis is on the theoretical implications.

1 Introduction

Mathematical and simulation modelling of Internet traffic, even at a single location, has proven to be a surprisingly complex task, which has been surrounded by substantial controversy. A simple view of the traffic, at any given point, is that it is an aggregation of "flows", where each flow is a set of packets with shared source and destination.

The first models for aggregated Internet traffic were based on standard queueing theory ideas, using the exponential distribution to model flow durations. These models have the advantage of being tractable for standard time series analysis. But a number of studies of Internet traffic have suggested that Internet flows often have heavy tailed duration distributions, and that the aggregated traffic exhibits long range dependence, see *e.g.* Paxson (1994), Garrett and Willinger (1994), Paxson and Floyd (1995) and Crovella and Bestavros (1996). 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 convenient conceptual view of this behavior is given in Figure 1. Individual flows through a link are represented as horizontal lines (which start at the time of the first packet, and end at the last). A random vertical height ("jittering", see *e.g.* pages 121-122 of Cleveland 1993) is used for convenient visual separation. Their vertical aggregation constitutes the full traffic passing through the link. The time durations (*i.e.*, lengths) of the flows shown in Figure 1 appear to follow a "heavy tailed" distribution, in that there are a few very long flows (sometimes termed "elephants"), and also many very small flows (sometimes termed "mice"). If these durations were exponentially distributed with the same mean, then there would be far more "medium size" flows, as shown in Figure 2 of Hannig, Marron, Samorodnitsky and Smith (2001). These elephants cause the aggregated flow to be long range dependent. In particular,

[*]Department of Computer Science, University of North Carolina at Chapel Hill, NC 27599-3175; fhernand,smithfd@cs.unc.edu

[†]Dept. of Statistics, Univ. of North Carolina at Chapel Hill, NC 27599-3260; marron@stat.unc.edu

[‡]Dept. of Statistical Science, Cornell Univ., Ithaca, New York 14853; gennady@orie.cornell.edu

[§]School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, New York 14853

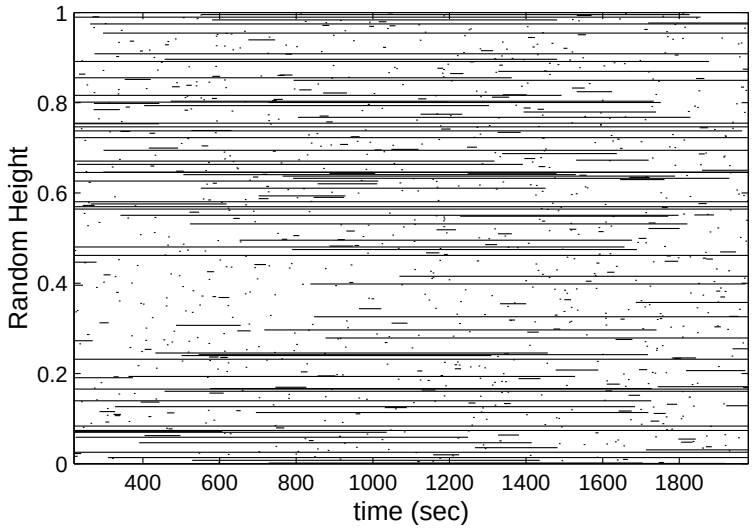

Figure 1: *Mice and elephants visualization of IP flows. Shows how heavy tail durations can lead to long range dependence of aggregated traffic.*

even at rather widely separated time points, there will be some common elephants, resulting in correlation between the total traffic at those time points. The above theory is a precise mathematical quantification of this concept.

The data shown in Figure 1 were gathered from IP (Internet Protocol) packet headers, during approximately 40 minutes 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, Hernández-Campos, Jeffay and Ott (2001). To eliminate visual boundary effects, only those flows which cross a time window of the central 80 % are considered here. There were 115548 such flows, and to avoid overplotting only a random sample of 1000 are shown in Figure 1.

While the above appealing framework of heavy tail duration distributions leading to long range dependence appears complete, more recent work has questioned both the heavy tail duration distributions, see Downey (2000), and Gong *et al.* (2001) and also the long range dependence, see Cao, Cleveland, Lin and Sun (2001). The controversy surrounding the first question is the main topic of the present paper. The second question has been resolved by appropriate visualization across a wide range of "scales" by Hannig, Marron and Riedi (2001).

Downey (2000) suggests that the light tailed log-normal distribution may give a better fit to many duration distributions than the heavy tailed Pareto. A naive view suggests that this is inconsistent with the above theory, because heavy tails appear to be critical. However, Hannig, Marron, Samorodnitsky and Smith (2001) show that contrary to previous notions, log-normal durations are *not* contradictory to long range dependence of the aggregated traffic.

Gong *et al.* (2001) present a number of important ideas on this topic. First, they point out that one can never completely determine "tail behavior" (in the classical asymptotic sense) of a distribution, based only on data. For example, each data set always has a largest data point, and the underlying distributional behavior beyond that point (and frequently anywhere within an order of magnitude or more of that point) cannot be reliably determined from the data. This concept motivates their important idea that distributional properties should really be investigated only over "appropriate ranges" of the data. In particular, any data set will contain very rich information in some regions (*e.g.* in the "main body" of the distribution), and very

sparse information in others (*e.g.* in the "tails").

A convincing and useful solution to the statistical problem of understanding the richness of distributional information from a set of data is provided in the companion paper, Hernández-Campos, *et al.* (2002b). Useful visual tools are applied to Internet traffic data in Section 2 (of that paper), which give a clear understanding of which distributional aspects are "important underlying structure", and which are "due to sampling variability". A data set whose size (number of flows well into the millions) is much larger than many of those that have appeared in published papers is analyzed. A naive view of such a large data set is "now we know the tail". But more careful consideration from the above perspective suggests that the only effect of a larger sample is that the region where we have a clear understanding of distributional properties becomes larger (but there is still a region of uncertainty far enough out in the tails).

A major result of the analysis of Section 2, of Hernández-Campos, *et al.* (2002b), is that the tail of the distribution has some strong "wobbles", of a type not present in the tails of classical distributions such as the log-normal or Pareto. It is tempting to attribute these wobbles to sampling variability. However, the statistical visualization suggests this is false. Deeper confirmation comes from repeating the analysis for a number of additional data sets. These not only exhibit the same amount of wobbles, but even *wobble exactly the same way in the same places*. This confirms the idea that these wobbles are important underlying distributional phenomena, and not sampling artifacts.

What causes the wobbles? This question is considered in Section 3, of Hernández-Campos, *et al.* (2002b). Several previously suggested distributional concepts are combined to find models which do fit the data (including wobbles) to the degree possible with the information at hand, in an intuitively meaningful way. In particular it is seen that mixtures of either 3 log-normals or else 3 double Pareto log-normals give an acceptable fit. From a classical asymptotic tail index viewpoint, these two distributions can be viewed as contradictory, since a mixture of log-normals is "light tailed" (in particular having all moments finite), while the fit double Pareto log-normal is "heavy tailed" (with an infinite variance, *i.e.*, second moment). This is another example of the interesting "distributional fragility" ideas raised by Gong *et al.* (2001), who made the very important observation that frequently a variety of models can give "good fit in the tails" (precisely because the distributional information is very sparse there). Based on the insights about variability that follow from our graphics, this is very consistent, and highlights the fact that one can never use data alone to distinguish between such models. Instead of debating which model is "right", it makes more sense to think about the "collection of models that are consistent", and what can be learned from them as a whole. Consequences which hold for all of the reasonable models then seem the most compelling.

A deep and important issue of this type is: What is the impact of these statistically significant wobbles in the tail of the duration distribution on the above elegant theory, suggesting that heavy tails of the duration distribution cause long range dependence? Downey (2001) provided interesting statistical evidence of these wobbles, through an analysis based on the concept of "tail index". The classical definition of "tail index", from extreme value theory (see *e.g.* Chapter 1 of Resnick (1987) for an introduction) is the asymptotic rate of decay of the (underlying theoretical) cumulative distribution function. Downey analyzes an empirical version of this, and shows that it often does not stabilize as one moves out in the tail (completely consistent with the "wobbliness" discussed above), and concludes that duration distributions are "not heavy tailed". He goes on to suggest that another cause needs to be found for the observed long range dependence in aggregated Internet traffic.

Another goal of the present paper is a deeper look at these issues, from the above viewpoint of "understanding tail behavior in various regions, with attention paid to sampling variability".

This motivates refining the notion of "tail" to cover three important cases. The part of the tail that is beyond the last data point (thus with no information at all in the data) is called the "extreme tail". The part of the tail where there is some data present, but not enough to reliably understand distributional properties is called the "far tail". The part of the tail where the distributional information in the data is "rich" is called the "moderate tail". These concepts are heuristic, so there is no sharp boundary, *e.g.*, between the far and moderate tail, but these concepts provide the needed framework for understanding the data analysis given in Section 2 of Hernández-Campos, *et al.* (2002b).

Another visualization, shown in Section 2, of the present paper is an "effective tail index plot", showing that, as observed by Downey (2001), indeed the tail index does not seem to stabilize as one moves farther out in the tail. This indeed suggests that the simplest models from extreme value theory are not applicable. But there is a richer tail index theory, based on the idea of "regular variation", which allows for wobbliness, as long as the wobbliness diminishes as one moves out in the tail. This theory is described, and related to the data in Section 3.1, where it is also seen that tail wobbles in the duration distribution "are somewhat smoothed" in the autocovariance of the aggregated traffic. This richer extreme value theory allows for wobbles in the tail of the duration distribution, and also implies long range dependence in the aggregated traffic (according to the conventional theory).

While a regular variation assumption is consistent with all that is observed in the data, it is not completely satisfying, because it leans on a type of asymptotics that involves eventual stabilization of the tail index (even if the stabilization is "slow"). Worse, it appears to be ultimately driven by far and extreme tail behavior, where the information in the data is unacceptably sparse. This motivates understanding other ways in which heavy tailed durations can lead to long range dependence. A key idea here comes from the observation that the "effective tail index" is "quite often" in a range which is classically associated with "yielding long range dependence". While Downey correctly found oscillation, it happens mostly in a range where *any of the observed tail indices would imply long range dependence* if they only stabilized. This motivates an alternate mathematical approach, where the tail index does not stabilize, but instead only stays within certain bounds (and perhaps does that only "most of the time"). The final important contribution of this paper is an enhanced theory which shows such assumptions can also yield long range dependence. In particular, in Section 3.2, very mild conditions (very broadly consistent with the data analyzed above) on the duration distribution are presented, which yield asymptotic behavior (polynomial decay of the autocovariance function) that is symptomatic of long range dependence.

The main contributions of our larger combined paper, Hernández-Campos, et. al. (2002a), from the perspective of the networking community are recapped in Section 4.

2 Duration distribution analysis

In this section HTTP responses, gathered from the UNC main link during April of 2001 are considered. "Flow" is now defined to be the set of packets associated with a single HTTP data transfer, and "flow duration" is the time between the first and last packets. To allow study of diurnal effects, packets were gathered over 21 four hour blocks, over each of the 7 days of the week, and for "morning" (8:00AM-12:00AM), "afternoon" (1:00PM-5:00PM) and "evening" (7:30PM-11:30PM) periods on each day. The total number of HTTP flows over the four hour blocks ranged from ~ 1 million (weekend mornings) to ~ 7 million (weekday afternoons). The HTTP duration distributions are analyzed separately for each of these 21 time blocks. The 21 analyses were surprisingly similar, so to save space, only the results for Thursday

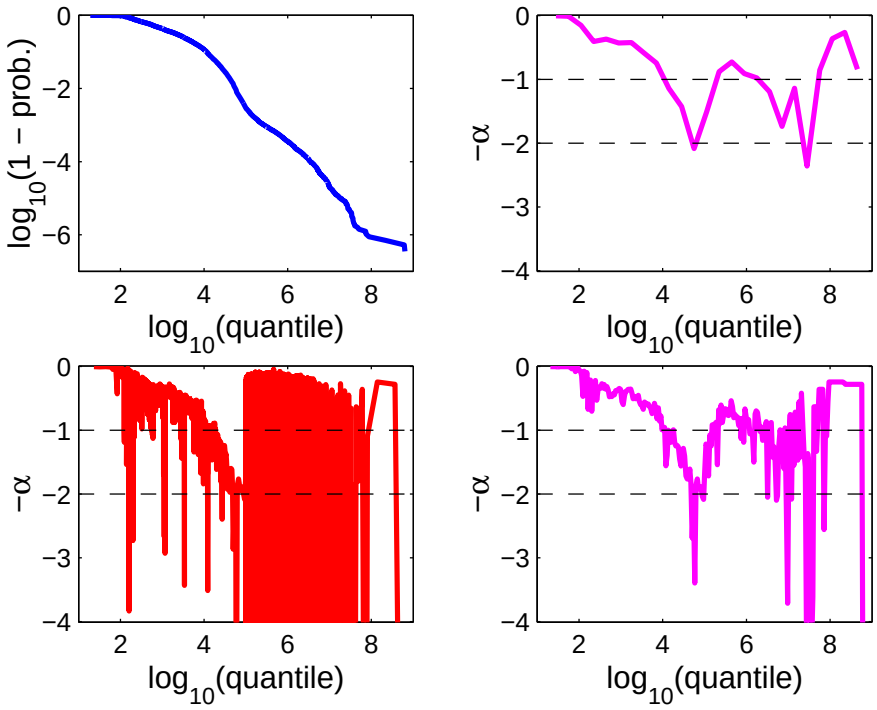

Figure 2: *Effective Tail Index plots for the Thursday morning HHTP response duration data. Shows that while the effective tail index does not stabilize, it still"mostly" stays in critical range* $\alpha \in (1, 2)$.

morning are shown for most purposes. However the other analyses can be conveniently viewed in files indicated below, in the web directory `http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails/`.

The striking similarity of the wobbles shown in Figure 4 of Hernández-Campos, *et al.* (2002b) suggests that it is worth trying to understand and perhaps to model them. This is done in Section 3 of that paper, where it is seen that mixtures, either of three log-normal or of three double Pareto log-normal distributions provide a good fit. The mixture components are then used to cast light on likely phenomena for generation of the wobbles.

As noted in Downey (2001) the log log CCDF is also very useful for understanding "effective tail index". In particular, the slope of the curve in Figure 3 of Hernández-Campos, *et al.* (2002b) can be taken as a notion of "effective tail index" (multiplied by -1). However, while "slope" seems visually clear in graphics like those shown in Figure 3 of that paper, it can be slippery to properly define and work with. Downey (2001) uses a type of numerical differentiation that gives some useful results, but care and numerical expertise are important here, as illustrated in Figure 2

The top left panel of Figure 2 shows the same blue curve as in Figure 3 of Hernández-Campos, *et al.* (2002b) (but with a different aspect ratio) for easy comparison. A simple way to find slopes is to take simple difference quotients, shown in red in the lower left panel of Figure 2. Difference quotients of noisy data can be very unstable, and a rather large number of them here are actually undefined because of zero denominators. Difference quotients with a zero denominator (and those smaller than -4) are mapped to -4 in this view. If one is willing to view the small difference quotients as "noise artifacts", then most attention should be paid to the large values. Interpreting the upper edge of the red region as minus the effective local tail index, α, note that there is an impression of "frequently $\alpha \in (0, 2)$". This range is quite noteworthy, because it is the classical tail index range of the duration distribution which results

in long range dependence (or even non-stationarity) of the aggregated traffic.

While the red difference quotient analysis shown in the lower left panel of Figure 2 is suggestive, it is far from conclusive, because it is not clear how many of the difference quotients are near the upper edge. A more convincing view can be obtained by more careful numerical differentiation of the blue curve in the top panel, which is roughly similar to smoothing the red curve. A simple approach is to use more widely spaced points for forming the difference quotients. This is done in the right panels in Figure 2, where difference quotients are computed based on an equally spaced grid with width δ. The values of δ were carefully chosen to illustrate the main ideas, which turned out to be $\delta = 0.3$ for the bottom panel and $\delta = 0.03$ for the bottom panel. In the bottom right panel, the purple version of the effective tail index, computed using the fairly small $\delta = 0.03$ smoothed away much of the variability present in the red version in the bottom left panel, however there is still substantial variability present, suggesting this δ is still small. The effective tail index for the smoother purple curve in the top right panel, using $\delta = 0.3$, now has the noise completely eliminated. The danger of too large a δ is that it will also smooth away clearly important and systematic changes in the slope, *i.e.*, the effective tail index. However, observe that the general shape of the effective tail index in the top panel is the same as that in the bottom panel (modulo the similar downward noise for the bottom panel), which shows that no strong distortion is present. Thus the purple curve in the top right panel does a good job of reflecting "effective tail index".

The first interesting feature of this effective tail index α is that it wobbles substantially, with the wobbles following those of the blue curve in the top left panel. In particular the purple curve is lower where the blue curve is steeper. The second interesting feature is that it is nearly always in the range $\alpha \in (0, 2)$, which as noted above is classically associated with long range dependence. Similar analyses for the other time blocks can be viewed in the file UNC2001RS1allCCDFSfullcombine.pdf, in the above web directory. The lessons are generally quite similar. An exception is that in many cases the purple $-\alpha$ curve did not dip below the threshold at 2, and Thursday morning was chosen for display, because this did happen then, thus motivating the generality of the theoretical work in Section 3.2.

This view shows clearly that while the effective tail index does not "stabilize" in any meaningful sense, at least over regions where we have useful distributional information, it seems very likely to generate long range dependence in the aggregated traffic. This idea is backed up in one way by the results in Section 7 of Gong *et al.* (2001). Another view, based on a deeper and more general theory is given in Section 3.2.

3 Improved long range dependence theory

This section explores several types of mathematical theory which are motivated by the above analysis. The wobbliness of the tails, visible for example in Figure 2, is seen to be consistent with the notion of "regular variation" in Section 3.1. Under this assumption, the wobbles must diminish as one moves far enough out in the tails, and thus the classical notion of tail index still holds. However, if this stabilization occurs, Figure 2 shows that it happens in a tail region where the distributional information in the data is sparse. This corresponds to the case where either the number of mixture components in the models of Section 3 of Hernández-Campos, *et al.* (2002b) does not grow, or else components farther out in the tail have a diminishing impact.

In the spirit of simultaneous consideration of several models that fit the data (and can't be reliably distinguished from the data alone), it makes sense to also consider mathematics where the tail need not be regularly varying. Figure 2 also shows that while the effective tail index does not stabilize, it is "usually" within a range that is associated with the generation of

long range dependence. The mathematics of Section 3.2 feature very mild assumptions on the effective tail index, as plotted in Figure 2, which will still result in long range dependence, in the sense of a polynomial decay of the autocovariance function. This corresponds to the case where the number of significant mixture components in the models of Section 3 of Hernández-Campos, *et al.* (2002b) continues to grow as one moves farther out in the tail.

For convenience of analysis, this section considers a deliberately simple mathematical model for data of the type illustrated in Figure 1. 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. Our model has been called a "fluid queue with Poisson input" and a "model with $M/G/\infty$ input". 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. Marron, Hernández-Campos and Smith (2001) have studied the effectiveness of this approximation, and suggested a richer model. The duration times (the random lengths of the line segments), are independent, identically distributed, with cumulative distribution function (CDF) $F(x)$ and complementary CDF (CCDF) $\overline{F}(x) = 1 - F(x)$. Aggregation of the traffic is represented by X_t, the number of active flows (line segments in Figure 1) at time t.

A common notion of long range dependence can be expressed in terms of the rate of decay of the autocovariance

$$R(t) = \mathrm{cov}(X_s, X_{t+s}).$$

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

$$R(x) = \lambda \int_x^\infty \overline{F}(y)\, dy \tag{1}$$

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

3.1 Varying slopes in classical heavy tail theory

This section studies the classical notion of "regularly varying tails", which allows wobbly tail behavior as seen in Figure 2.

A common notion of "heavy tailed distribution" is that $\overline{F}(x) \sim x^{-\alpha}$, in the sense that for some $C > 0$ and $\alpha > 1$,

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{Cx^{-\alpha}} = 1.$$

In this case, a is called the "tail index".

It is usual also to refer to such distributions as having power, or Pareto, tails. Such distributions are really a particular case of distributions with "regularly varying tails", which also cause long range dependence. As defined in, for example Section 0.4.1 of Resnick (1987), a distribution is said to be regularly varying at ∞, with exponent $-\alpha$ when for every $x > 0$

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha}.$$

As seen in Karamata's Theorem, see Section 0.4.2 of Resnick (1987), a useful characterization of a regularly varying tail is the existence of functions $\varepsilon(t)$ and $c(x)$, where for $x \geq 1$,

$$\overline{F}(x) = c(x)x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(t)}{t} dt\right)$$

and where $\varepsilon(t) \to 0$ as $t \to \infty$ and $c(x) \to c \in (0, \infty)$ as $x \to \infty$.

The function $\varepsilon(t)$ has a very direct connection to the wobbles observed in the tails of the log-log CCDF in Figure 2. We will assume for simplicity that $c(x) \equiv 1$. This guarantees existence of a probability density, $f(x)$. Note that

$$\begin{aligned} f(x) &= F'(x) = \left(1 - \overline{F}(x)\right)' \\ &= \alpha x^{-(\alpha+1)} \exp\left(\int_1^x \frac{\varepsilon(t)}{t} dt\right) - x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(t)}{t} dt\right) \frac{\varepsilon(x)}{x} \\ &= x^{-(\alpha+1)} \exp\left(\int_1^x \frac{\varepsilon(t)}{t} dt\right) (\alpha - \varepsilon(x)) \\ &= \frac{\overline{F}(x)}{x} (\alpha - \varepsilon(x)). \end{aligned} \tag{2}$$

This shows that for the function $\varepsilon(x)$ to result in a probability distribution $F(x)$, it must be restricted to $\varepsilon(x) \leq \alpha$, for all $x \geq 1$. Figure 2 is an empirical version of a plot of $\overline{F}(x)$ as a function of x, on the log - log scale. The slopes in this plot, whose empirical versions are studied in Figure 2, are essentially the derivative

$$\frac{d}{dy} \log \overline{F}(e^y) = -\frac{f(e^y) e^y}{\overline{F}(e^y)} = -(\alpha - \varepsilon(e^y)),$$

a very simple function of the $\varepsilon(x)$ from Karamata's Theorem.

This framework shows that the "regularly varying" functions allow a large degree of tail "wobbling", such as seen in Figure 2. A simple example is

$$\varepsilon(x) = \frac{\sin x}{x^\beta}, \tag{3}$$

for $x \geq 1$ and $\beta > 0$, which will result in wobbles of the magnitude observed there.

While the log-log CCDF of the distribution can and does wobble considerably, it is perhaps worth noting that under the above model, the resulting aggregated traffic autocovariance $R(t)$, tends to be "smoother" because of the integration in (1). In particular, when the duration distribution $F(x)$ is regularly varying,

$$R(t) = \int_t^\infty \overline{F}(x)\, dx \sim \frac{1}{\alpha - 1} t \overline{F}(t),$$

as $t \to \infty$, which may oscillate much less. The rest of this section is devoted to making this idea precise.

Assume for example that

$$|\varepsilon(t)| = O\left(\frac{1}{t}\right),$$

as $t \to \infty$, and that

$$\int_x^\infty \frac{\varepsilon(t)}{t} dt = O\left(\frac{1}{x^2}\right),$$

as $x \to \infty$. An example of this is (3) with $\beta = 1$. Then the distribution has a Pareto tail in the sense that $\overline{F}(x) \sim c x^{-\alpha}$, as $x \to \infty$. This results in a classical symptom of heavy tail dependence of the aggregated traffic: $R(t) \sim t^{-(\alpha-1)}$.

By (2)

$$\alpha - \frac{x f(x)}{\overline{F}(x)} = \varepsilon(x) = O\left(\frac{1}{x}\right),$$

as $x \to \infty$, and so

$$\left| \frac{\overline{F}(x)}{xf(x)} - \frac{1}{\alpha} \right| \sim \varepsilon(x) = O\left(\frac{1}{x}\right).$$

Now the above calculations, together with

$$t\overline{F}(t) = t^{-(\alpha-1)} \exp\left(\int_1^t \frac{\varepsilon(u)}{u} du\right) = (\alpha-1)\int_t^\infty x^{-\alpha} dx \exp\left(\int_1^t \frac{\varepsilon(u)}{u} du\right)$$

give

$$\begin{aligned}
\left| \frac{R(t)}{t\overline{F}(t)} - \frac{1}{(\alpha-1)} \right| &= \frac{\left| (\alpha-1)\int_t^\infty x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(u)}{u} du\right) dx - (\alpha-1)\int_t^\infty x^{-\alpha} \exp\left(\int_1^t \frac{\varepsilon(u)}{u} du\right) dx \right|}{(\alpha-1)\, t\overline{F}(t)} \\
&\leq \frac{\int_t^\infty x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(u)}{u} du\right) \left| 1 - \exp\left(-\int_t^x \frac{\varepsilon(u)}{u} du\right) \right| dx}{t\overline{F}(t)} \\
&\leq (1+o(1)) \frac{\int_t^\infty x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(u)}{u} du\right) \left| \int_t^x \frac{\varepsilon(u)}{u} du \right| dx}{t\overline{F}(t)} \\
&\leq (1+o(1))\, t^{-2} \frac{\int_t^\infty x^{-\alpha} \exp\left(\int_1^x \frac{\varepsilon(u)}{u} du\right) dx}{t\overline{F}(t)} \\
&\sim \frac{1}{1-\alpha} t^{-2},
\end{aligned}$$

as $t \to \infty$. Thus while the slope of the log-log duration CCDF converges to α at the slow rate x^{-1}, the slope of the resulting aggregated autocovariance (in the same log-log scale) converges to $\alpha - 1$ at the much faster rate t^{-2}.

3.2 Varying slopes give long range dependence

This section considers the case where the tail wobbles visible in Figure 5 may not be of regularly varying type, in the sense that the effective tail index does not stabilize as one moves farther out in the tail of the distribution.

The main result is that under suitable mild assumptions, allowing behavior of the type observed in Figure 2, one has behavior that is symptomatic of long range dependence, in the sense that

$$R(x) \geq k x^{-(\alpha-1)}, \tag{4}$$

for some constant $k > 0$.

Assume that for some $c > 0$, and $\alpha > 1$,

$$\overline{F}(x) \geq c x^{-\alpha},$$

for $x \in I_n = (a_n, b_n)$, $n = 1, 2, \cdots$, with $a_1 < b_1 < a_2 < b_2 < \cdots$, satisfying for some $M > 1$,

$$\frac{a_{n+1}}{b_n} \leq M, \frac{b_n}{a_n} \geq 1 + \frac{1}{M},\ n = 1, 2, ...$$

This structure is intended to quantify the notion that the effective tail index is "usually but not always" smaller than α, in particular allowing the purple curves in Figure 2 to occasionally fall

below the level -2. For $x > 0$, find the "indices that bracket x by a_n":

$$\begin{aligned}\eta_+(x) &= \min\{j : a_j \geq x\},\\ \eta_-(x) &= \max\{j : a_j < x\} = \eta_+(x) - 1.\end{aligned}$$

Note that $a_{\eta_-(x)} < x \leq a_{\eta_+(x)}$.

Now we check that these assumptions give (4). Assume first that $x \in \left[b_{\eta_-(x)}, a_{\eta_+(x)}\right]$. Then

$$\frac{a_{\eta_+(x)}}{x} \leq \frac{a_{\eta_+(x)}}{b_{\eta_-(x)}} = \frac{a_{\eta_+(x)}}{b_{\eta_+(x)-1}} \leq M. \tag{5}$$

Hence

$$\begin{aligned}R(x) &= \int_x^\infty \overline{F}(y)\,dy \geq \int_{a_{\eta_+(x)}}^{b_{\eta_+(x)}} cy^{-\alpha}dy \qquad (6)\\ &= \frac{c}{\alpha-1}\left(a_{\eta_+(x)}^{-(\alpha-1)} - b_{\eta_+(x)}^{-(\alpha-1)}\right)\\ &= \frac{c}{\alpha-1}a_{\eta_+(x)}^{-(\alpha-1)}\left(1 - \left(\frac{b_{\eta_+(x)}}{a_{\eta_+(x)}}\right)^{-(\alpha-1)}\right)\\ &\geq \frac{c}{\alpha-1}\left(1 - \left(\frac{M}{M+1}\right)^{\alpha-1}\right)a_{\eta_+(x)}^{-(\alpha-1)}\\ &\geq \frac{c}{\alpha-1}M^{-(\alpha-1)}\left(1 - \left(\frac{M}{M+1}\right)^{\alpha-1}\right)x^{-(\alpha-1)}\\ &\geq \frac{c}{\alpha-1}\left(M^{-(\alpha-1)} - (M+1)^{-(\alpha-1)}\right)x^{-(\alpha-1)}.\end{aligned}$$

Next assume that $x \in \left[a_{\eta_-(x)}, b_{\eta_-(x)}\right]$. If $b_{\eta_-(x)} - x \geq x$, then

$$\begin{aligned}R(x) &= \int_x^\infty \overline{F}(y)\,dy \geq \int_x^{b_{\eta_-(x)}} cy^{-\alpha}dy \qquad (7)\\ &\geq c\int_x^{2x} y^{-\alpha}dy = \frac{c\left(1 - 2^{-(\alpha-1)}\right)}{\alpha-1}x^{-(\alpha-1)}.\end{aligned}$$

Finally, if $b_{\eta_-(x)} - x < x$, then as in (5)

$$\frac{a_{\eta_+(x)}}{x} \leq \frac{a_{\eta_+(x)}}{b_{\eta_-(x)}/2} \leq 2M,$$

from which it follows that, as for (6),

$$R(x) \geq \frac{c}{\alpha-1}2^{-(\alpha-1)}\left(M^{-(\alpha-1)} - (M+1)^{-(\alpha-1)}\right)x^{-(\alpha-1)}. \tag{8}$$

The bound (4) follows from (6), (7) and (8).

4 Conclusions

The larger paper, Hernández-Campos, *et al.* (2002a), from which this was drawn made two major contributions of interest to the networking community.

The first contribution (appearing elsewhere as Hernández-Campos, *et al.* (2002a) to meet the page requirements of this proceedings) was the presentation of a number of useful techniques for the study of heavy tailed distributions in network modelling. The concepts of "extreme", "far" and "moderate" tail regions facilitate understanding of how sampling variation affects this modelling. Simulation, combined with appropriate graphical display, is useful for identification of these regions. Mixture models provide a natural method for finding interpretable subpopulations. Mixtures of 3 double Pareto log-normals accurately model HTTP response sizes.

The second contribution (appearing here) was the generalization of the "classical" theory of heavy tail durations leading to long range dependence, in a well motivated and relevant direction. The data analysis suggested that a serious gap in the relevance of the classical theory is the assumption of a fixed tail index (central to the usual definition of "heavy tailed"). This problem was overcome using the more realistic concept of "variable tail index", and a more general theory was established in which this improved notion of "heavy tailed" was shown to still lead to long range dependence (in terms of polynomial decay of the autocovariance function).

5 Acknowledgement

The collaboration of this paper is a result of the course OR778 at Cornell University, during the Fall of 2001. The research of J. S. Marron was supported by NSF Grant DMS-9971649, and of Gennady Samorodnitsky by NSF grant DMS-0071073. The terminology of "moderate", "far" and "extreme" tails arose from discussion with D. Towsley. The authors would like to thank NLANR MOAT and the WAND research group for making packet header traces publicly available. We especially thank Joerg Micheel for his help.

References

[1] Cao, J., Cleveland, W. S., Lin, D. and Sun, D. X. (2001) The effect of statistical multiplexing on Internet packet traffic: theory and empirical study. Internet available at: `http://cm.bell-labs.com/cm/ms/departments/sia/InternetTraffic/webpapers.html`.

[2] Cleveland, W. S. (1993) *Visualizing Data*, Hobart Press, Summit, New Jersey, U.S.A.

[3] Cox, D. R. (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.

[4] Crovella, M. E. and A. Bestavros, A. (1996) Self-similarity in world wide web traffic evidence and possible causes, *Proc. of the ACM SIGMETRICS 96*, pages 160–169, Philadelphia, PA.

[5] Downey, A. B. (2001) The structural cause of file size distributions, Proc. of IEEE/ACM MASCOTS'01.

[6] Downey, A. B. (2001) Evidence for long tailed distributions in the internet, ACM SIGCOMM Internet Measurement Workshop, November 2001.

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

[8] Gong, W., Liu, Y., Misra, V. and Towsley, D. (2001) On the tails of web file size distributions, *Proc. of 39-th Allerton Conference on Communication, Control, and Computing*. Oct. 2001. Internet available at: `http://www-net.cs.umass.edu/networks/publications.html`.

[9] Hannig, J., Marron, J. S. and Riedi, R. (2001) Zooming statistics: Inference across scales, *Journal of the Korean Statistical Society*, 30, 327-345.

[10] Hannig, J., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2001) Log-normal durations can give long range dependence, unpublished manuscript, web available at `ftp://ftp.orie.cornell.edu/pub/techreps/TR1320.ps`.

[11] 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.

[12] Hernández-Campos, F., Marron, J. S., Samorodniztky, G., and Smith, F. D. (2002a) Variable Heavy Tailed Durations in Internet Traffic. Internet available at `http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails`.

[13] Hernández-Campos, F., Marron, J. S., Samorodniztky, G., and Smith, F. D. (2002b) Variable Heavy Tailed Durations in Internet Traffic, Part I: Understanding Heavy Tails. In Proc. of IEEE/ACM MASCOTS'02.

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

[15] Marron, J. S., Hernández-Campos, F. and Smith, F. D. (2001) A SiZer analysis of IP Flow start times, unpublished manuscript. Internet available at `ftp://ftp.orie.cornell.edu/pub/techreps/TR1333.pdf`.

[16] Paxson, V. (1994) Empirically-Derived Analytic Models of Wide-Area TCP, Connections. *IEEE/ACM Trans. on Networking*, 2, 316–336.

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

[18] Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*, Springer-Verlag, New York.

[19] 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.

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

[21] 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.