Variable Heavy Tailed Durations in Internet Traffic

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

Abstract

This paper 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.

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 IP 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

[*]Dept. of Computer Science, Univ. of North Carolina at Chapel Hill, NC 27599-3175; fhernand,smithfd@cs.unc.edu
[†]Department of Statistics, University of North Carolina at Chapel Hill, NC 27599-3260; marron@stat.unc.edu
[‡]Department of Statistical Science, Cornell University, Ithaca, New York 14853; gennady@orie.cornell.edu
[§]School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853

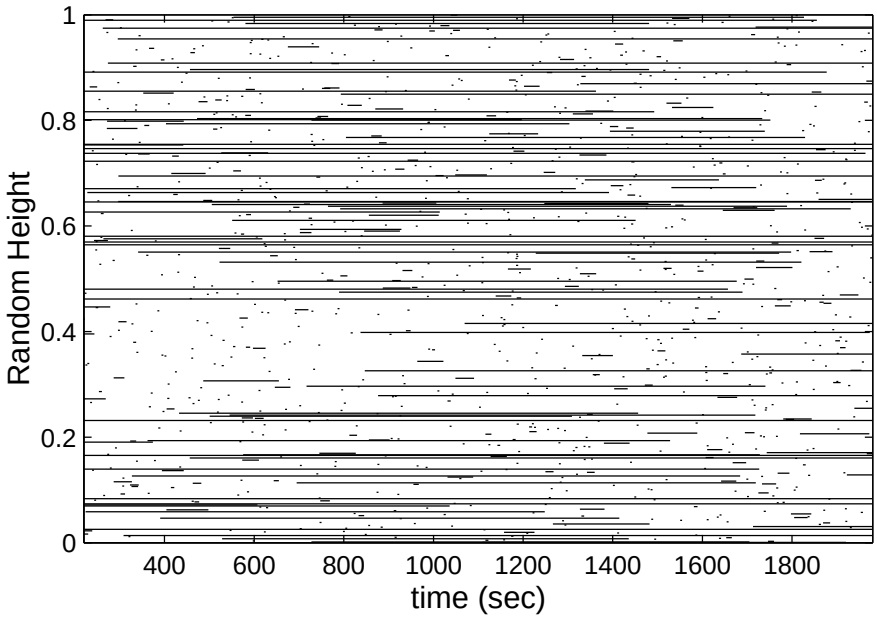

Figure 1: *Mice and elephants visualization of IP flows. Shows how heavy tail durations can lead to long range dependence of aggregated traffic.*

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 short 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, 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 depen-

dence appears complete, more recent work has questioned both the heavy tail duration distributions, see Downey (2000), and Gong, Liu, Misra and Towsley (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, Liu, Misra and Towsley (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 the first major goal of this paper. Useful visual tools are applied to Internet traffic data in Section 2, 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 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. 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, Liu, Misra and Towsley (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.

Another visualization in Section 2 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 4.1, 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 4.2, 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 this paper, from the perspective of the networking community are recapped in Section 5.

2 Duration distribution analysis

In this section a different data set from that of Figure 1 is analyzed. This time 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 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/`.

2.1 Pareto tail fitting

Figure 2 shows how well the Thursday morning HTTP duration distribution (based on $n = 5,663,605$ data points) is fit by the standard Pareto distribution. The visual device used here is called a Q-Q plot, because it allows graphical comparison of the quantiles of a "theoretical Pareto distribution" with the quantiles of the data set. In particular the red curve is constructed by plotting theoretical quantiles on the horizontal axis against the sorted data values on the vertical axis (on a log-log scale, to avoid a few large values dominating the picture). If the data quantiles were the same as the theoretical quantiles (this should approximately happen when the fit is "good"), the red curve should follow the 45 degree line, shown as green in Figure 2. See Fisher (1983) for a good overview of Q-Q plots, and a variety of related statistical tools.

For better insight into which part of each distribution is represented by which part of the red curve, labelled plus signs are shown for some selected quantiles. One reflection of the heavy tail nature of these data is the fact that the 0.99 quantile (only 1 percent of the data are larger than this) appears near the middle of this display. This shows that there are very few "elephants" (the data on the upper right), and a very large number of "mice" (the bulk of the data on the lower left). The particular Pareto distribution shown here was chosen by quantile matching. In particular the two Pareto parameters α and σ were chosen to make the theoretical and empirical 0.8 and 0.99 quantiles (shown as small circles) the same. Thus the red curve crosses the green line at these quantiles.

The Pareto distribution, i.e. the closeness of the red line to the green curve, might be deemed "acceptable". There is some "wobbling", which one might expect to be due to the natural sampling variability. On the other hand, the sample size is quite large, so maybe the amount of wobbling is statistically significantly greater than could be expected from truly Pareto data. The blue curves provide a visual device for simple understanding of this issue. They are an overlay of 100 simulations of data sets of the same size, $n = 5,663,605$, from the same Pareto distribution. If the data were truly Pareto, then the wobbles of the red curve would lie mostly within the blue envelope. This is roughly true for the very largest data values, but generally the wobbles veer far outside of the blue envelope (which for much of the range of the data is so close to the green line that it disappears underneath), showing this difference is statistically significant, and thus not due to the natural sampling variation. A clear conclusion is that the Pareto distribution is

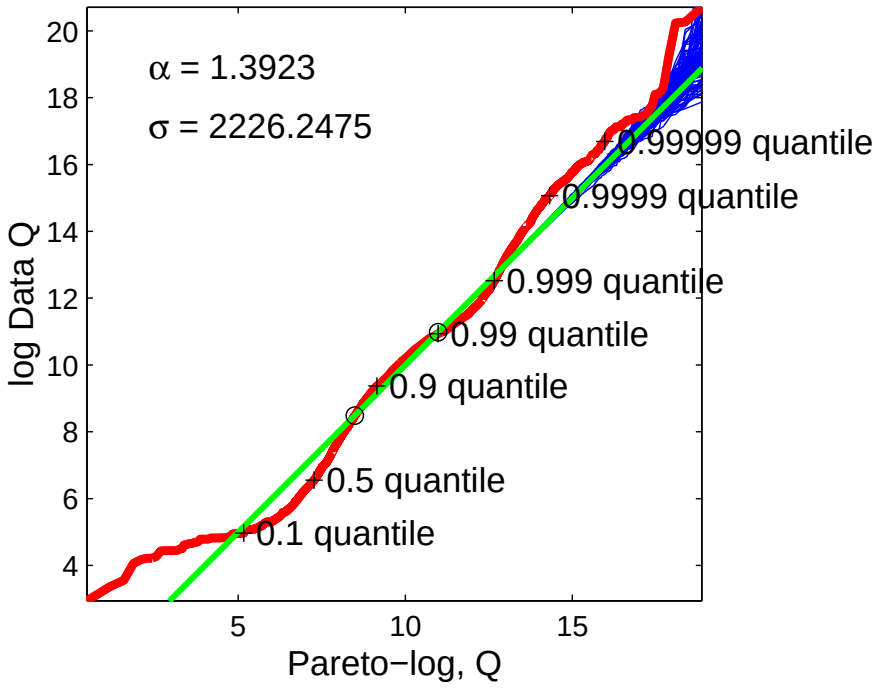

Figure 2: *Pareto Q-Q plot (red) for the Thursday morning HTTP response duration data. Compare to 45 degree line (green) and simulated versions (blue).*

not a precise fit to these data (not surprising with a sample so large). A similar analysis, with very similar conclusions, of all 21 time blocks is available in the file UNC2001RS1allQQparcombine.pdf in the above web directory.

In addition to allowing conclusions of the above type, the visualization in Figure 2 also begins to provide an answer to the question: where do the data provide clear distributional information? The information is clearly very strong (in the sense that the blue envelope is completely underneath the green curve) up to nearly the 0.9999 quantile (the point where only 0.01% of the data are larger). This region includes both the "body of the distribution", and the "moderate tail". Note that this includes HTTP responses of all sizes up to about 1.2 megabytes (perhaps the term "elephants" can be used for responses that are larger than this, among the collection of HTTP traffic), and there are about 560 of these among the 5.6 million total responses. For the top 500 responses distributional information is understandably sparser, but the blue envelope in Figure 2 suggests that some useful insights may still be available, even up to about the 0.99999 quantile (where only the top 50 data values lie). This region is termed the "far tail" of this distribution. Finally the "extreme tail" is the region larger than the biggest data point (the right end of the red curve), 980 megabytes for this data set.

Downey (2000) suggests that the log normal fit may be expected to be better. A similar analysis to

Figure 2 was performed, with the log-normal replacing the Pareto, but the results seemed slightly worse. In particular, in addition to the tail wobbliness observed here, there is also substantial curvature away from the green line. Such a picture is not included here, because it is tangential to the main points of this paper, but full results can be viewed in the file UNC2001RS1allQQlncombine.pdf in the above web directory.

2.2 Variable Tail Index

A strength of the Q–Q visualization shown in Figure 2 is that it allows precise comparison to a given distribution, coupled with immediate understanding of the sampling variability (shown by the blue envelope), and thus of the moderate, far and extreme tails. A weakness of the Q-Q visualization is that it can only be constructed in the context of a particular theoretical distribution. An obvious choice for the theoretical distribution may not be available, especially if one would like to model the "tail wobbles" apparent in Figure 2.

A common alternate visualization of tail behavior in data, which has the advantage of not being tied to any theoretical distribution, is the log log Complementary Cumulative Distribution Function (CCDF) plot, shown in Figure 3, for the same data as in Figure 2. In this view, the sorted data values (called "empirical quantiles" in Section 2.1, and appearing on the vertical axis in Figure 2) are plotted on the horizontal axis, while the corresponding CCDF (simply an equally spaced grid, from 0 to 1) is plotted on the vertical axis.

If the data came from a Pareto distribution, the blue curve in Figure 3 would be nearly linear, and the slope of the line would be the Pareto shape parameter (also called "tail index") α. Again for clarity as to where the data lie in this plot, some selected quantiles are indicated. Matching these with the corresponding quantiles in Figure 2 shows an interesting correspondence. In particular, the wobbles in Figure 2 correspond directly with the wobbles in Figure 3.

A serious weakness of the graphic in Figure 3 is that it shows nothing about the important underlying statistical variability, and thus provides no indication of the boundary between the moderate and far tails. It is natural to suspect that the wobbles are just artifacts of the sampling process, and can be ignored. However, the deep analysis of Figure 2 suggests that these wobbles are systematic, not random, variation.

A similar analysis, for all 21 time blocks is available in the file UNC2001RS1allCCDFfullcombine.pdf in the same web directory. This file may be the most interesting of those posted, because it is rather surprising how *similar* all 21 of these curves look. In particular they lie nearly on top of each other, over a surprisingly large range of the data.

Another view of this is given in Figure 4, where the same log log CCDF plot is shown, for all 21 four hour time blocks, as an overlay. Different colors are used to indicated different days, with Sunday being plotted last with cyan (light blue). Note that for the first 99% of the data range (recall the 0.99 quantile appears

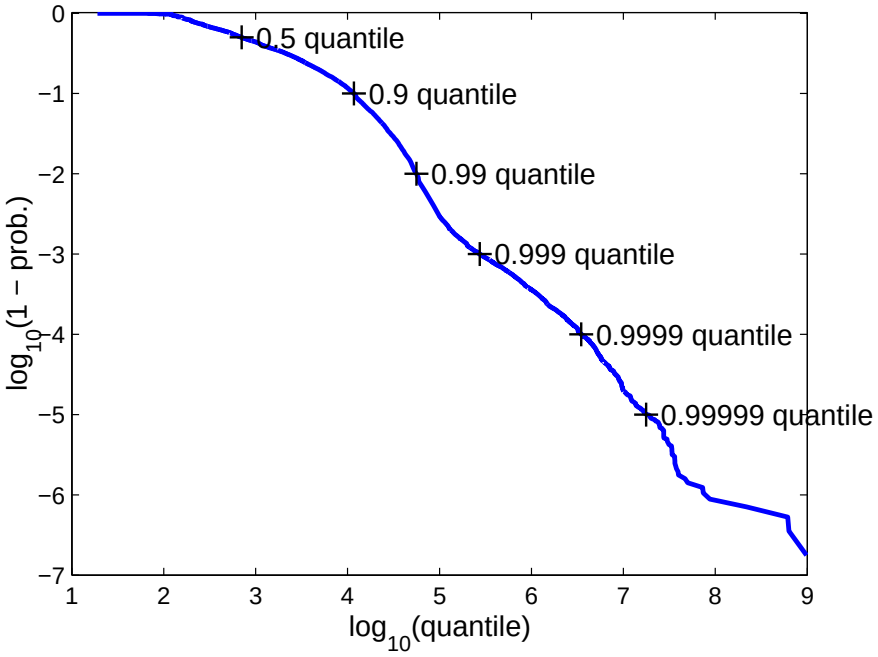

Figure 3: *Log log CCDF plot for the Thursday morning HTTP response duration data. Shows wobbly tail, inconsistent with most standard distributions.*

as -2 on the vertical axis), cyan is nearly the only color visible, because the other 18 curves are so close to these. An exception is one red curve, which is substantially different because of the very unusual presence of more than 480,000 (out of about 2.1 million total) responses of size exactly 381 bytes. We are unsure of the cause of this but a deeper look revealed that a very large percentage of all the responses with that size were sent by a single UNC host to a single server. Hence we suspect this may be due to a malfunction, a hostile action, or a non-HTTP use of port 80, over a substantial period of time (the Friday morning and afternoon time blocks also had an unusually large number of responses of exactly this size).

The similarity of the curves in Figure 4 provides a very different confirmation of the lesson learned from Figure 2: the wobbles in the tail are systematic, *not* due to sampling variability. This time the variability is studied by replicating the experiment over some different time blocks. One goal of this study was to understand diurnal (i.e. time of day and day of week) effects. Such effects have a large impact on total traffic and system usage, driven by easily understandable differences in user behavior. We expected this obviously differing user behavior to also have a major impact on response size distributions (e.g. during peak times, more "business" web browsing, with students and faculty looking for educational resources, staff browsing e-commerce sites etc., with more multimedia rich recreational browsing being done at off peak times) Thus we found the constancy of distribution over time blocks quite surprising.

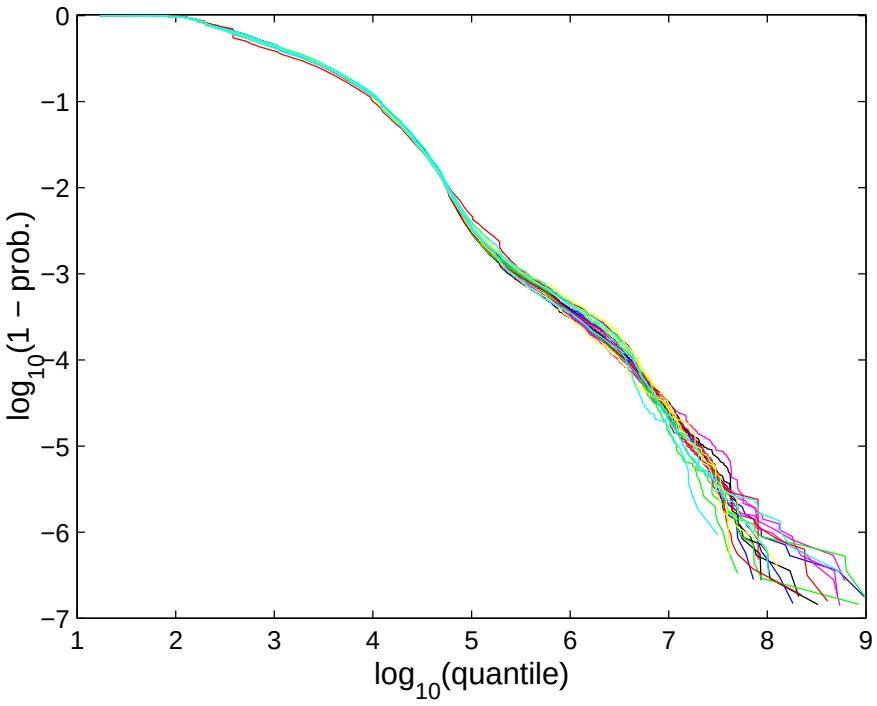

Figure 4: *Log log CCDF plots for HTTP response duration data for all 21 time blocks. Note very similar pattern, showing "wobbles" are not sampling artifacts.*

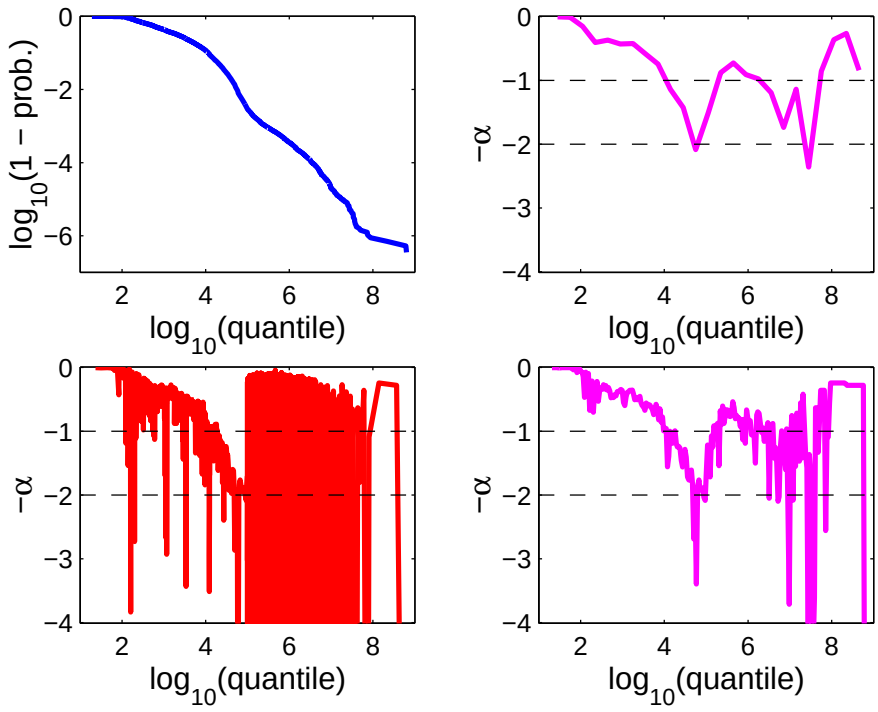

Figure 5: *Effective Tail Index plots for the Thursday morning HTTP response duration data. Shows that while the effective tail index does not stabilize, it still"mostly" stays in critical range* $\alpha \in (1, 2)$.

The striking similarity of the wobbles shown in Figure 4 suggests that it is worth trying to understand and perhaps to model them. This is done in Section 3, 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 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, 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 5.

The top left panel of Figure 5 shows the same blue curve as in Figure 3 (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 5. 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 5 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 5, 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 4.2.

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, Liu, Misra and Towsley (2001). Another view, based on a deeper and more general theory is given in Section 4.2.

3 Improved distribution modelling

Figures 2 and 4 provide a strong suggestion that the wobbles in the tail of the distribution represent important underlying structure. In this section, that structure is modeled, which provides a vehicle for potential explanations. Section 3 of Gong, Liu, Misra and Towsley (2001) contains a good overview of possible mechanisms for generation of duration distributions of the type observed above.

Downey (2000) presented some attractive arguments for why distributions of file sizes could be expected to be log-normal. The main idea is that most files are modifications of other files, and that such modifications are often effectively viewed as "multiplicative changes" in the file size. Aggregation of a sequence of independent changes of this type may result in a multiplicative central limit theorem, thus yielding a log-normal distribution. While Downey was working explicitly with file sizes, such mechanisms seem to be at play with response size distributions as well.

Reed (2001) presents the double Pareto log-normal distribution, which is the product of a double Pareto random variable (having density proportional to $x^{-\alpha-1}$ for $x > 1$ and proportional to $x^{-\beta-1}$ for $x < 1$) with an independent lognormal random variable. This distribution can be viewed as extending Downey's ideas by incorporating an independent exponential number of random shocks. Allowing the number of multiplicative shocks to be random not only seems a little more realistic, it has the large advantage of yielding a Pareto-like polynomial tail of the distribution. This feature is quite interesting, especially in view of Figure 2, where it is seen that the Pareto gives a fit to the actual response size distribution that is not completely unreasonable.

Figure 6 assesses the goodness of fit of the double Pareto log-normal distribution, to the data shown in Figures 2 and 3. This time the view is again the log log CCDF, so the blue curve is the same as in Figure 3. The red curve shows the log log CCDF for a double Pareto log-normal distribution with parameters chosen for good visual impression. Some attempts at maximum likelihood estimation failed, perhaps because the parameters are nearly not identifiable (observed during the visual fitting process), or because there are multiple local solutions generated by the wobbles. To visually reflect the level of sampling variability, once again 100 simulated data sets, also of the same size $n = 5,663,605$, were drawn, and the resulting log log CCDFs are also plotted, this time in purple. In the same spirit as Figure 2, the purple envelope gives easy visual insight into the separation between the moderate and far tails, i.e. where the distributional information in the data is rich, and where it is sparse.

The red curve in Figure 6 is nearly linear over much of its range, showing that its tail corresponds closely to that of a Pareto (which is exactly linear). This property is not shared by the log-normal, although it *can* hold approximately over a quite wide range of quantiles, which drives the results of Hannig, Marron, Samorodnitsky and Smith (2001). This asymptotic, i. e. extreme tail, convergence of the double Pareto log-normal log log CCDF to linear may be a conceptual advantage over the conventional log-normal.

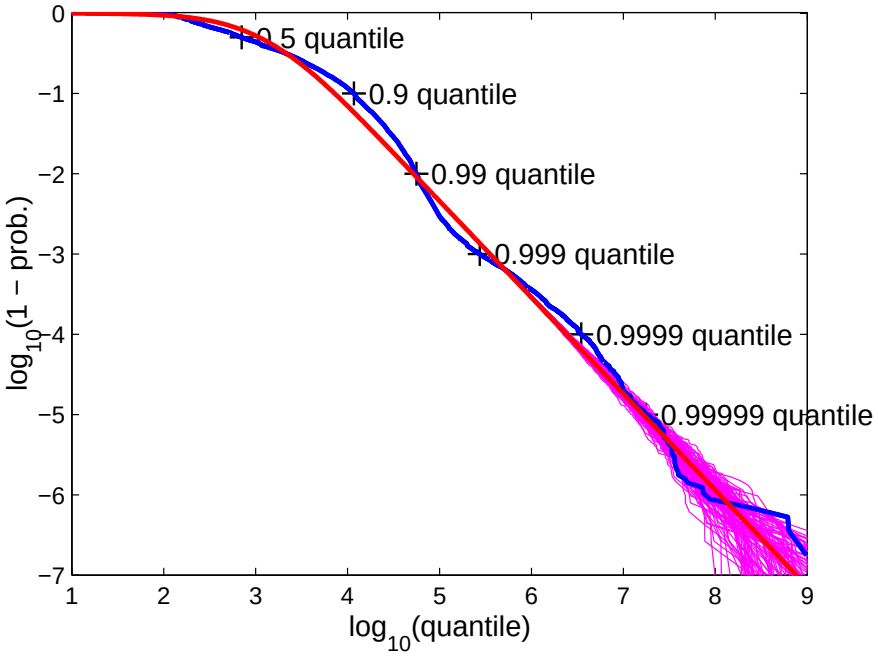

Figure 6: *Log log CCDF plot for the Thursday morning HTTP response duration data, together with a visually fit double Pareto log-normal.*

The purple envelope in Figure 6 shows that while the double Pareto log-normal seems to head globally in the right direction, it is still far from a "good fit" (which happens when the blue curve lies mostly in the purple envelope). As observed in Section 2, all of the departures are caused by wobbles in the tail of the distribution of the response size data, and happen in the moderate tail.

The distributions considered so far do not have the flexibility to capture the wobbles, because their tails are inherently smooth. While there are many ways to generate probability distributions with wobbly tails (e.g. by using "piece-wise" approaches), the most intuitively appealing is mixture modeling. Mixture models arise very naturally in the context of a population that is composed of several subpopulations. Wobbles of the type observed above result when these subpopulations have very different distributions.

Figure 7 shows the result of fitting a mixture of three double Pareto log-normal distributions to the same response size data set as above. The format is the same as in Figure 6. A mixture of two was able to explain a large share of the wobbliness, see the file UNC2001RS1CCDFdpln2msim10.ps in the above web directory, but not all, so the mixture of three is shown here. As noted above, double Pareto log-normal parameters even for a single population do not appear to be straightforward to estimate. This problem becomes far more challenging for mixture models, where estimation is notoriously slippery, even for mixtures of simple distributions. Hence the parameters of the red fit have again been tuned for good visual impression (through

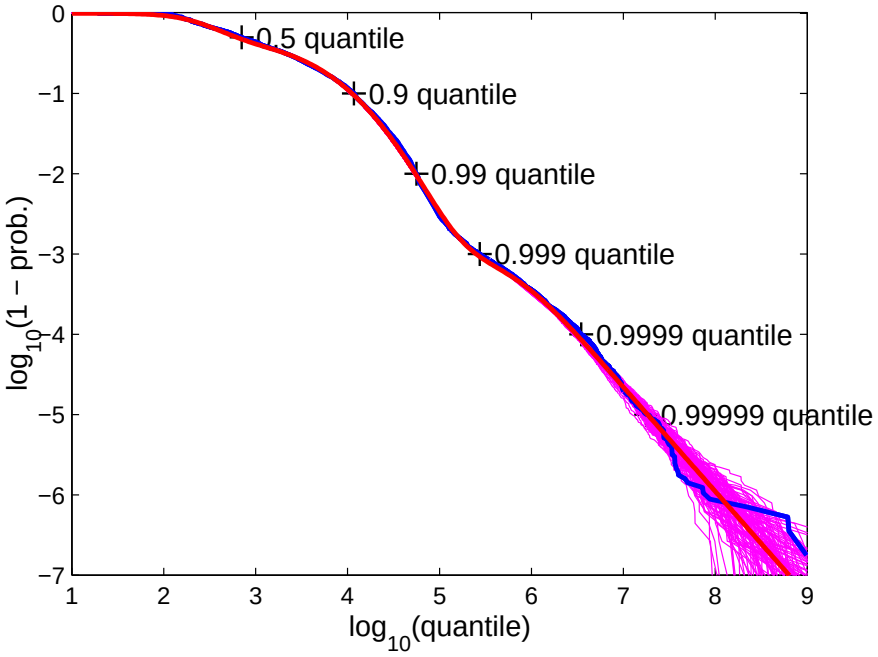

Figure 7: *Log log CCDF plot for the Thursday morning HTTP response duration data, together with a visually fit mixture of 3 double Pareto log-normals.*

a painstaking trial and error process).

The red fit curve in Figure 7 lies nearly completely on top of the blue curve showing the log log CCDF of the data. The only substantial departure is on the lower right, the far tail, where the purple envelope reveals that the variability is mostly well within that expected from the sampling process.

While the fit in Figure 7 looks impressively good (especially for such a large sample size), it is important to resist the urge to "conclude that these data come from this model". First off, it must be kept in mind that the family of mixture distributions is extremely broad, and if enough components are included, almost any distribution can be well approximated. For example, the visual device of the purple envelope steers one away from the temptation to add a fourth mixture component to "explain" another wiggle beyond the 0.99999 quantile. This could be done, but it would be gross "overfitting", because that wiggle can not be separated from the random sampling noise. Second off, it is important to recall the nice "distributional fragility" ideas of Gong, Liu, Misra and Towsley (2001), and the idea that there are likely to be a family of different distributions that fit.

This point is made in Figure 8, which is the same as Figure 7, except that now a mixture of three log-normals is fit to the data. As above, a mixture of two log normals was attempted, but was not satisfactory, see the file UNC2001RS1CCDFln2msim10.ps in the above web directory. Again the population parameters

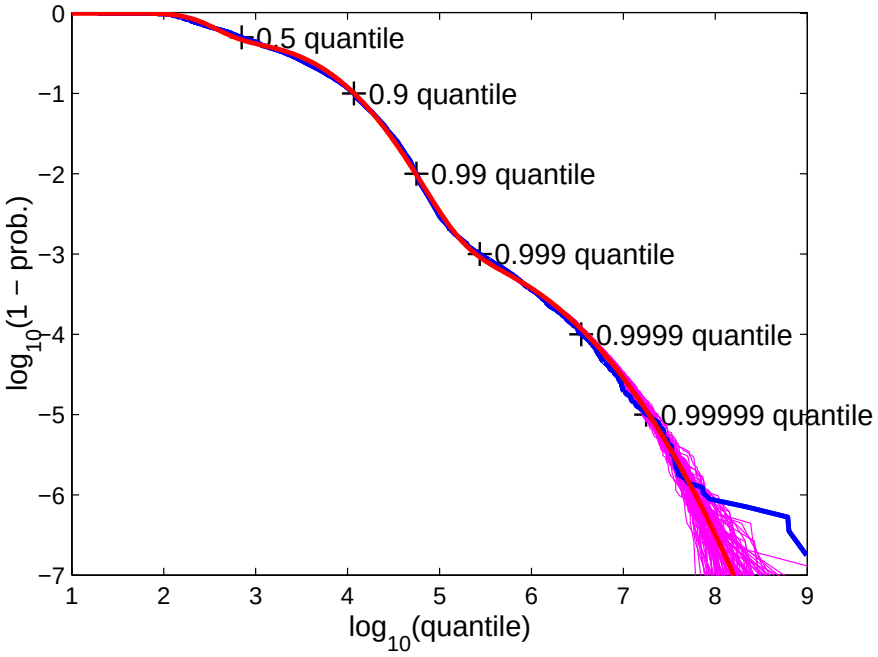

Figure 8: *Log log CCDF plot for the Thursday morning HTTP response duration data, together with a visually fit mixture of 3 log-normals.*

were fit by painstaking visual trial and error.

The fit in Figure 8 is again impressively good. A minor exception is for the extreme observations in the right part of the far tail, where the blue curve leaves the purple envelope. This suggests that the log-normal tail is not exactly right (which makes intuitive sense when comparing Downey's conceptual model with Reed's), but it is very close, and could clearly be captured by adding just one more mixture component.

As noted above, these two distributions are quite different in terms of classical asymptotic tail behavior. But the important point is that they are very similar in the moderate tail, and thus can not be distinguished using only the data. Hence, several models should be kept in mind for later analysis, and for simulation.

Models which fit as well as those shown in Figures 7 and 8, should be able to cast new insights in to the phenomena at hand. This calls for careful consideration of the chosen parameters. For the log-normal mixture in Figure 8, the parameters are the mean μ_i and standard deviation σ_i of the components (of the data in the natural log scale), and the component "weights" w_i (i.e. the probability a data point comes from that subpopulation). The numerical values, using i to index the component subpopulations, appear in the

table

i	1	2	3
μ_i	5.7	8.45	13.05
σ_i	0.6	1.2	1.55
w_i	0.55	0.4488	0.0012

.

The first two parts together include nearly 99.99% of the data, thus modelling the body of the data as shown in Figure 8. The slight wobble near the median appears because about 55% are in the first subpopulation, which is centered at $\log_{10}(\exp(5.7)) = 2.47$ on the scale of Figure 8, and almost all the rest are in the second, centered at $\log_{10}(\exp(7.45)) = 3.24$. The convexity near the 0.999 quantile is created by the interaction of these two components, with the third component, centered at $\log_{10}(\exp(13.05)) = 5.67$.

These parameters allow a simple and appealing explanation of the sub-populations. About 55% of the HTTP responses come for a population with sizes in the neighborhood of an order of magnitude of 10^2 bytes, which could be tiny layout images, small HTML pages (such as error status pages), and navigation bars in multi-frame pages. Most of the rest of the traffic has sizes with order of magnitude in the neighborhood of 10^4 bytes, which perhaps includes most standard HTML text pages and images. But there is a significant sub-population of far larger sizes, with sizes roughly in the neighborhood of 10^6, that perhaps are software, multimedia content (such as movies) and PDF documents.

We speculate that within each subpopulation, Downey's ideas of multiplicative averaging are indeed generating distributions similar to the log-normal. But the full distribution does not look log-normal, because there is not so much averaging occurring that it can bridge the large gap between these sub-populations.

The double Pareto log–normal distributions have more parameters. Here the notation of Reed (2001) is used. The distribution has the appealing simple representation of being a product of a mixture of Pareto random variables and a log-normal. The log normal mean parameter for each mixture component is ν_i and the standard deviation is τ_i. The Pareto factor has tail parameters α_i and β_i. Explicit numerical values for the red curve in Figure 7 are

i	1	2	3
ν_i	5.7	8.45	13.35
τ_i	0.6	1.2	0.75
α_i	2	10	1.3
β_i	2	10	1.3
w_i	0.6	0.399	0.001

Note that many of the parameters are surprisingly similar to the corresponding log-normal parameters given above. This is because when the tail parameters α_i and β_i are large, the Pareto mixture factor is close to 1, and thus negligible, so the distribution is nearly log-normal. The exception is the third component,

where the log-normal standard deviation τ_3 is substantially smaller, since "sub-population spreading" is accomplished by the very influential tail parameter $\beta_i = 1.3$. This tail parameter makes this distribution "heavy tailed" in the classical asymptotic sense.

Again there is a strong suggestion that Reed's ideas of population generation are working on these subpopulations, but these three are separated by too many orders of magnitudes for their differences to be averaged out.

Interesting possibilities for future work include a more careful identification of the subpopulations, and a study of how they evolve over time. In particular both the subpopulation parameters, and also their relative weights (the w_i) probably change over time in ways that are worth study. Also new subpopulations are likely to appear in the future. Finally it would be of keen interest to extend this type of analysis to other types of TCP traffic (only HTTP is studied here), which would likely include other interesting sub-populations, such as file-sharing applications.

3.1 Other Data Sources

An important question about the modelling discoveries made above is: how well do they generalize? In particular, we have only taken a deep look at HTTP response sizes from the UNC main link, and these population properties could be artifacts of only that location.

To study this issue we have applied a similar analysis to more HTTP response size data sets, derived from the archives of the National Laboratory for Applied Network Research (`http://www.nlanr.net/`).

We first analyzed traces from the University of Auckland, and found quite similar structure. The trace collection consists of seven 24-hour long header traces taken at the Internet access link of the University of Auckland in mid April, 2001. We derived a response duration data set following the same procedure we developed for the UNC traces. Graphics are not shown here to save space. But analogs of Figures 2, 3, 5, 6 and 7 can be found in files starting with NZ2001... in the above web directory. The lessons from these follow the same train of thought as above. No single distribution provides an acceptable fit, but a mixture of three double Pareto log-Normal distributions gives an excellent fit, using somewhat different parameters.

To investigate whether the main points also extend beyond universities, we next analyzed data from the New Zealand Internet Exchange (NZIX). At the time of the traffic capture, NZIX served as a peering point for six telecommunication companies. The traces comprise 6 days of packet headers collected in July 2000. Again the lessons were very similar, so it is not worth showing the full analysis. However, once again analogs of Figures 2, 3, 5, 6 and 7 can be found in files starting with NZIX00... in the above web directory. This time, the most important result, the good fit of a mixture of three double Pareto log-Normal distributions, is shown in Figure 9. Here $n = 857,172$ HTTP responses were found in a four hour period between 8:00AM

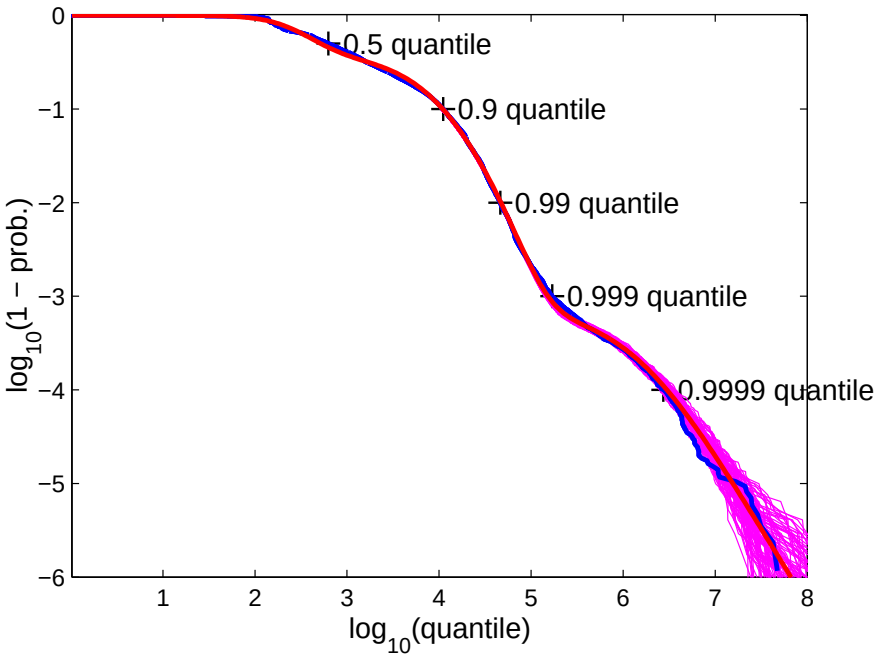

Figure 9: *Log log CCDF plot for the Thursday morning HTTP response duration data, together with a visually fit mixture of 3 double Pareto log-normals, for data from the New Zealand Internet Exchange.*

and 12:00Noon, during April of 2000.

The fit is of similar high quality as that shown for the UNC data in Figures 7 and 8. Hence, the main ideas of this paper appear to carry over very well to other contexts.

Another example, with very similar main lessons, is shown in Figure 10. This time the data come from the University of Auckland, in New Zealand. The data are the durations of 610,816 HTTP responses from 1:00PM to 5:00PM, on Tuesday, April 4, 2001.

These new data, are again visually fit with a mixture of three double Pareto log Normal distributions. As in Figures 7, 8 and 9, the fit is exceptionably good, with the red and blue curves essentially coinciding over the moderate tail, and acceptable variation in the far tail (where the purple envelope "begins to fan out"). We suggest that these general lessons apply well beyond the UNC data that have driven most of our work.

4 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 3, is seen to be consistent with the notion of "regular variation" in Section 4.1. Under this assumption, the wobbles must diminish as one moves far enough out in

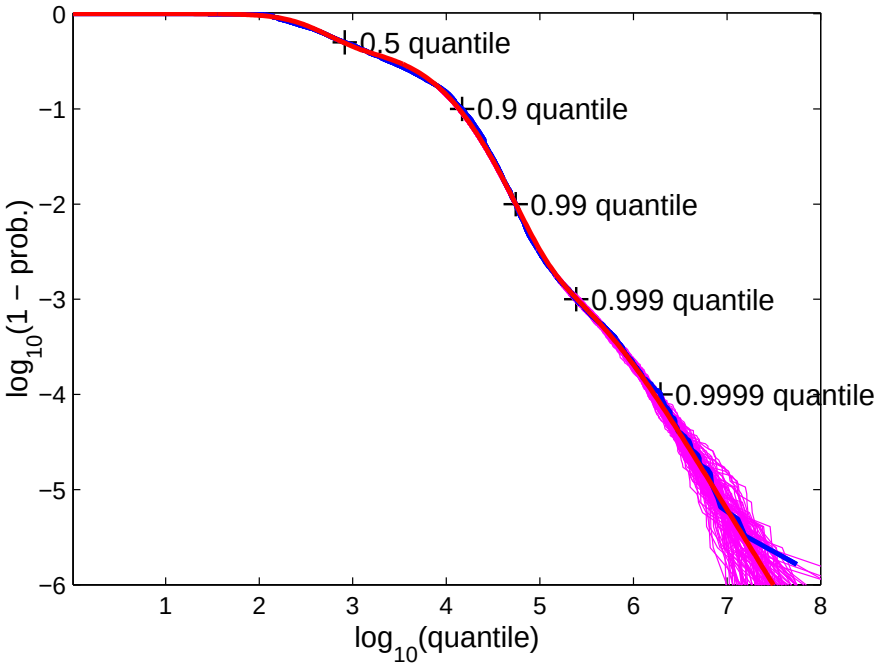

Figure 10: *Log log CCDF plot for the Tuesday afternoon HTTP response duration data, together with a visually fit mixture of 3 double Pareto log-normals, for data from the University of Auckland.*

the tails, and thus the classical notion of tail index still holds. However, if this stabilization occurs, Figure 5 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 does not grow, or else components farther out in the tail have a diminishing impact.

In the above 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 5 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 4.2 feature very mild assumptions on the effective tail index, as plotted in Figure 5, 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 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).

4.1 Varying slopes in classical heavy tail theory

This section studies the classical notion of "regular varying tails", which allows wobbly tail behavior as seen in Figure 3.

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 regular 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 3 above. 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 3 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 4, 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 "regular varying" functions allow a large degree of tail "wobbling", such as seen in Figure 3. 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)

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

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) convergence to $\alpha - 1$ at the much faster rate t^{-2}.

4.2 Varying slopes give long range dependence

This section considers the case where the tail wobbles visible in Figure 3 may not be of regular 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 4, one has behavior that is symptomatic of long range dependence, in the sense that

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

for some constant $k > 0$.

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

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

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 5 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

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

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

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

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).

5 Conclusions

This paper has made two major contributions of interest to the networking community.

First, a number of useful techniques are presented for the study of heavy tailed distributions in network modeling. 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.

Second, the "classical" theory of heavy tail durations leading to long range dependence is generalized in a well motivated and relevant direction. The data analysis suggests 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 is overcome using the more realistic concept of "variable tail index", and a more general theory is established in which this improved notion of "heavy tailed" is shown to still lead to long range dependence (in terms of polynomial decay of the autocovariance function).

6 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 Internet 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 Bestavros, A. (1996) Self-similarity in world wide web traffic evidence and possible causes, *Proceedings of the ACM SIGMETRICS 96*, pages 160–169, Philadelphia, PA.

[5] Downey, A. B. (2000) The structural cause of file size distributions, Proc. of IEEE/ACM MASCOTS'01, 2001.

[6] Downey, A. B. (2001) Evidence for long tailed distributions in the internet, ACM SIGCOMM Internet Measurement Workshop, November 2001. Internet available at `http://rocky.wellesley.edu/downey/longtail/`.

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

[8] 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.

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

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

[11] 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`.

[12] 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.

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

[14] 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`.

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

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

[17] Reed, W. J. (2001) The double Pareto - lognormal distribution - a new parametric model for size distributions, unpublished manuscript, Internet available at `http://www.math.uvic.ca/faculty/reed/`.

[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", *Proceedings 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.