Variable Heavy Tailed Durations in Internet Traffic, Part I: Understanding Heavy Tails

F. Hernández-Campos [1] J. S. Marron [2,3] G. Samorodnitsky [3,4] F. D. Smith [1]
[1] *Dept. of Computer Science, Univ. of North Carolina at Chapel Hill, NC 27599-3175*
[2] *Dept. of Statistics, Univ. of North Carolina at Chapel Hill, NC 27599-3260*
[3] *School of Operations Research and Industrial Eng., Cornell Univ., Ithaca, NY 14853*
[4] *Department of Statistical Science, Cornell Univ., Ithaca, NY 14853*
{fhernand,smithfd}@cs.unc.edu marron@stat.unc.edu gennady@orie.cornell.edu

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 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. The emphasis here is on understanding heavy tails.

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.* [16, 8, 17, 4]. An elegant mathematical theory, see *e.g.* [15, 3, 21, 12], 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 [2]) 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 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 [11]. 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 [20]. To eliminate visual boundary effects, only those flows which cross a time window of the central 80 % are considered here. There were 115,548 such flows, and to avoid overplotting, only a random sample of 1,000 is shown

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 [6, 9], and also the long range dependence, see [1]. 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 [10].

Downey suggests in [6] that the light tailed log-normal

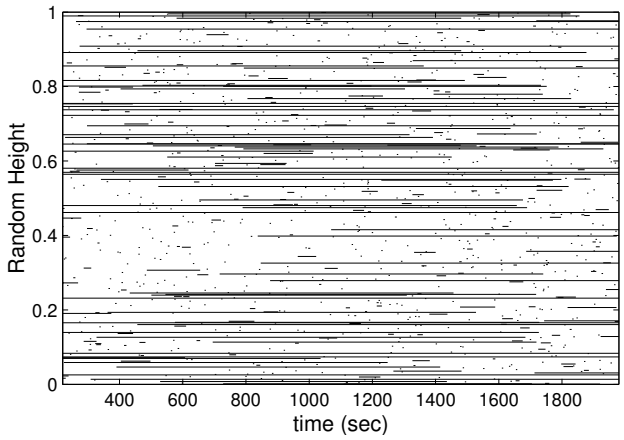

Figure 1. Mice and elephants visualization of IP flows. Shows how heavy tail durations can lead to long range dependence of aggregated traffic.

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, Hanning *et al.* show in [11] that contrary to previous notions, log-normal durations are *not* contradictory to long range dependence.

Gong *et al.* present in [9] 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 *et al.* who made the very important observation in [9] 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 provided in [5] 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 [19] 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, but they provide the needed framework for understanding the analysis in Section 2.

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 (seven days, three different 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 page `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 solid 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 solid curve should follow the 45 degree line (dashed). See [7] 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 solid curve, labelled plus

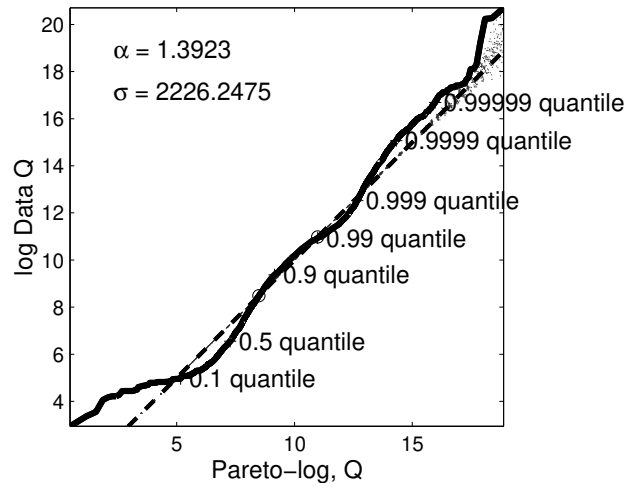

Figure 2. Pareto Q-Q plot (solid) for the Thursday morning response duration data. Compare to 45 degree line (dashed) and simulations (dotted).

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 solid curve crosses the dashed line at these quantiles.

The Pareto distribution, *i.e.*, the closeness of the solid line to the dashed 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 dotted 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 solid curve would lie mostly within the dotted envelope. This is roughly true for the very largest data values, but generally the wobbles veer far outside of the dotted envelope (which for much of the range of the data is so close to the dashed 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 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

above web page.

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 dotted envelope is completely underneath the dashed 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 dotted 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 solid curve), 980 megabytes for this data set.

Downey suggests in [6] 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 dashed 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 at the web page.

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 dotted 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 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 (an equally spaced grid, from 0 to 1) is plotted on the vertical axis.

If the data came from a Pareto distribution, the curve in Figure 3 would be nearly linear, and the slope of the line

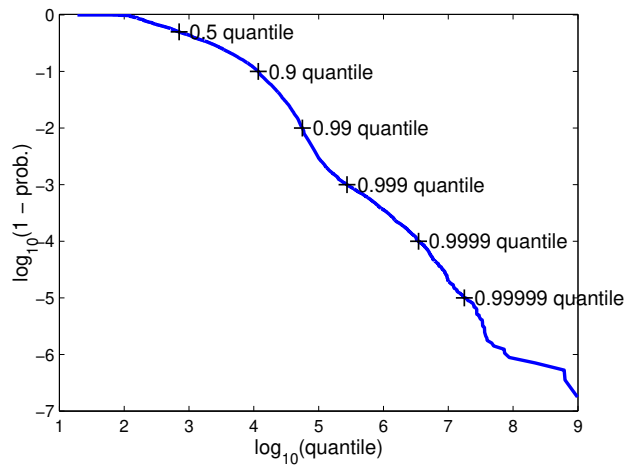

Figure 3. Log log CCDF plot for the Thursday morning response duration data. Shows wobbly tail, inconsistent with most standard distributions.

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 at the same web page 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. The similarity of the curves in this figure 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

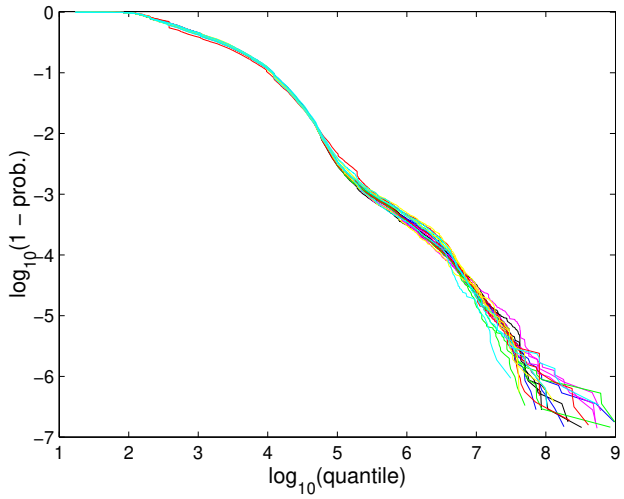

Figure 4. Log log CCDF plots for response duration data for all 21 time blocks. Note very similar pattern, showing "wobbles" are not sampling artifacts.

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.

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.

In the second part of the companion paper, [14], it is seen that, as noted in [5] 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). See Section 3 of [14] for more details and some deep theoretical implications.

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 [9] contains a good overview of possible mechanisms for generation of duration distributions of the type observed above.

Downey presented in [6] 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 presents in [18] 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 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 5 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 solid curve is the same as in Figure 3. The dashed 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. In the same spirit as Figure 2, the dotted 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 dashed curve in Figure 5 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 [11]. 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 log-normal.

The dotted envelope in Figure 5 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 solid curve lies mostly in the dotted 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 prob-

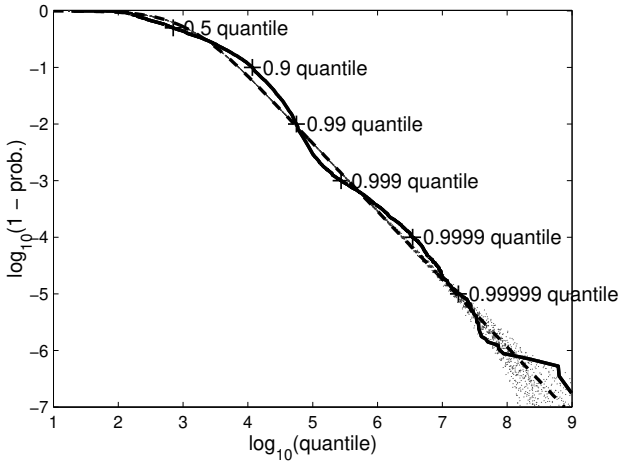

Figure 5. Log log CCDF plot for the Thursday morning HHTP response duration data, together a visually fit double Pareto log-normal (dashed curve).

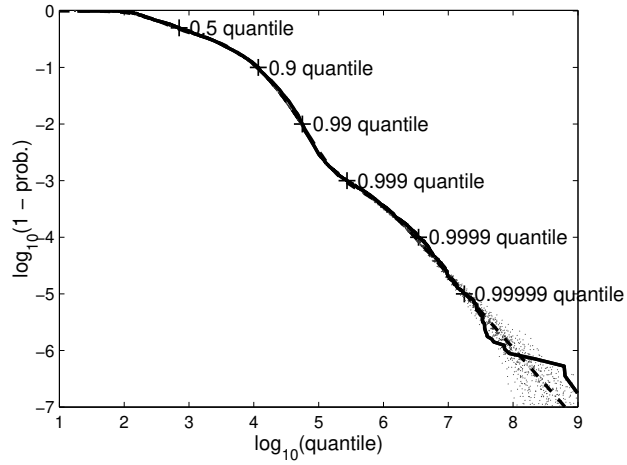

Figure 6. Log log CCDF plot for the Thursday morning response duration data, together a visually fit mixture of 3 double Pareto log-normals.

ability distributions with wobbly tails (*e.g.*, by using "piecewise" approaches), the most intuitively appealing is mixture modelling. 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 6 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 5. A mixture of two was able to explain a large share of the wobbliness, but not all, so the mixture of three is shown here (see web page). 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 dashed fit have again been tuned for good visual impression (through a painstaking trial and error process). The fit is excellent, and the only substantial departure is on the lower right, the far tail, where the dotted envelope reveals that the variability is mostly well within that expected from the sampling process.

While the fit in Figure 6 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 dotted 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 "distributional fragility" ideas of [9], and that there are likely to be a family of different distributions that fit.

This point is made in Figure 7, which is the same as Figure 6, 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 web page). Again the population parameters were fit by trial and error.

The fit in Figure 7 is again impressively good. A minor exception is for the extreme observations in the right part of the far tail, where the solid curve leaves the dotted 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 6 and 7, should be able to cast new insights in to the phenomena at hand. This calls for careful consideration of the chosen parameters. The larger version of this paper, [13], reports on the numerical values of the parameters of each of the three log-normal distributions that are mixed to fit the data in Figure 7. These parameters allow a simple and appealing explanation of the subpopulations. 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 and 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 mag-

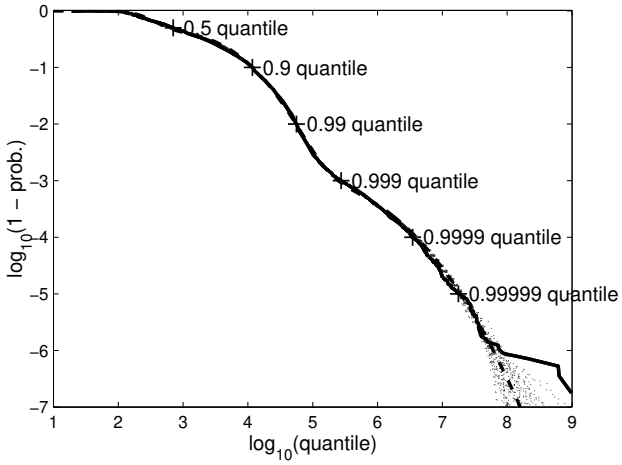

Figure 7. Log log CCDF plot for the Thursday morning response duration data, together a visually fit mixture of 3 log-normals.

nitude in the neighborhood of 10^4 bytes, which perhaps includes most standard HTML text pages and images. But there is a significant subpopulation of far larger sizes, with sizes roughly in the neighborhood of 10^6, that perhaps are software, multimedia content 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 could bridge the subpopulation gap.

The double Pareto log–normal distributions have more parameters. The numerical values of the parameters used in Figure 6 are given in [13]. Many of the parameters are surprisingly similar to the corresponding log-normal parameters given above. This is because large tail parameters, make the Pareto mixture factor close to 1, and thus negligible, so the distribution is nearly log-normal. 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. 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 subpopulations, 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 they can be found in the above web page. 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 here (see web page for this). This time, the most important result, the good fit of a mixture of three double Pareto log-Normal distributions, is shown in Figure 8. Here $n = 857,172$ HTTP responses were found in a four hour period between 8 AM and noon, during April of 2000. The fit is of similar high quality as that shown for the UNC data in Figures 6 and 7. Hence, the main ideas of this paper appear to carry over very well to other contexts.

4. Conclusions

The larger paper, [13], from which this was drawn made two major contributions of interest to the networking community.

The first contribution (detailed here) 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 elsewhere as [14], to meet the page requirements of these proceedings) was the generalization of the "classical" theory of heavy tail dura-

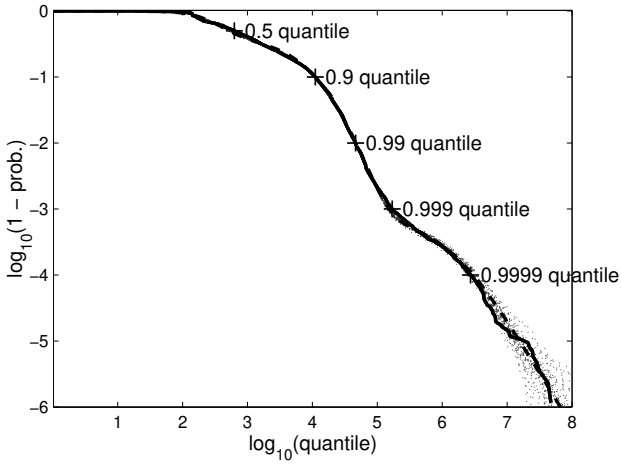

Figure 8. Log log CCDF plot for the Thursday morning response duration data, together a visually fit mixture of 3 double Pareto log-normals, for data from the New Zealand Internet Exchange.

tions 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] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. The effect of statistical multiplexing on Internet packet traffic, 2001.

[2] W. S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[3] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal, Proceedings 50th Anniversary Conference*, pages 55–74. The Iowa State University Press, 1984.

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

[5] A. B. Downey. Evidence for long tailed distributions in the Internet. In *Proc. of the ACM SIGCOMM Internet Measurement Workshop 2001*, November 2001.

[6] A. B. Downey. The structural cause of file size distributions. In *Proc. of IEEE/ACM MASCOTS'01*, 2001.

[7] N. I. Fisher. Graphical methods in nonparametric statistics: A review and annotated bibliography. *International Statistical Review*, 51:25–58, 1983.

[8] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar video traffic. In *Proc. of the ACM SIGCOMM '94*, pages 269–280, London, UK, 1994.

[9] W. Gong, Y. Liu, V. Misra, and D. Towsley. On the tails of web file size distributions. In *Proc. of 39-th Allerton Conference on Communication, Control, and Computing*, 2001.

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

[11] J. Hannig, J. S. Marron, G. Samorodnitsky, and F. D. Smith. Log-normal durations can give long range dependence. `ftp://ftp.orie.cornell.edu/pub/techreps/TR1320.ps`, 2001.

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

[13] F. Hernández-Campos, J. S. Marron, G. Samorodniztky, and F. D. Smith. Variable heavy tailed durations in Internet traffic. `http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails`, 2002.

[14] F. Hernández-Campos, J. S. Marron, G. Samorodniztky, and F. D. Smith. Variable heavy tailed durations in Internet traffic, part II: Theoretical implications. `http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails`, in submission., 2002.

[15] B. B. Mandelbrot. Long-run linearity, locally gaussian processes, H-spectra and infinite variance. *International Economic Review*, 10:82–113, 1969.

[16] V. Paxson. Empirically-derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking*, 2:316–336, 1994.

[17] V. Paxson and S. Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.

[18] W. J. Reed. The double pareto - lognormal distribution - a new parametric model for size distributions. `http://www.math.uvic.ca/faculty/reed/`, 2001.

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

[20] F. D. Smith, F. Hernández-Campos, K. Jeffay, and D. Ott. What TCP/IP protocol headers can tell us about the web. In *Proc. of ACM SIGMETRICS 2001/Performance 2001*, pages 245–256, Cambridge, MA, June 2001.

[21] M. Taqqu and J. Levy. Using renewal processes to generate LRD and high variability. In E. Eberlein and M. Taqqu, editors, *Progress in probability and statistics*, pages 73–89. Birkhaeuser, Boston, 1986.