Visualization Challenges in Internet Traffic Research

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

October 5, 2002

Abstract

This is an overview of some recent research, and of some open problems in the visualization of internet traffic data. One challenge comes from the sheer scale of the data, where millions (and far more if desired) of observations are frequently available. Another challenge comes from ubiquitous heavy tail distributions, which render standard ideas such as "random sampling will give a representative sample" obsolete. One more challenge is the visual representation of (and even the definition of) "common constant transfer rates" in a large scatterplot.

1 Introduction and background

The area of internet traffic measurement and modelling has a pressing need for novel and creative visualization ideas. The issues and the data are both complex, yet few researchers in that area (with some notable exceptions) are aware of the power of visualization for addressing the problems, and understanding complicated behavior.

The internet shares some similarities to the telephone network. Both are gigantic, worldwide networks for the transmission of information. Both share the notion of "connection", generally between two points. For this reason the first models for internet traffic were based on standard queueing theory, with assumption of Poisson arrival of connections, and exponentially distributed times of connection duration.

A large body of exciting work during the 1990's revealed that these assumptions were grossly inadequate, and far different models were usually much more appropriate. In particular, duration distributions were seen to exhibit heavy tails (caused by both far shorter, and also far longer connections than typically found in telephone traffic), and time series of aggregated traffic exhibit bursty behavior and long range dependence. An elegant mathematical theory, demonstrating how heavy tail durations can lead to long range dependence was developed by Mandelbrot (1969), Cox, D. R. (1984), Taqqu and Levy (1986), Leland, Taqqu, Willinger and Wilson (1994), Crovella and Bestavros (1996), Heath, Resnick and Samorodnitsky (1998), and Resnick and Samorodnitsky (1999).

This theory is deep and compelling, and gives good description of observed behavior in a wide range of circumstances. However, there has been recent controversy at several points.

One controversial point has been the issue of the heaviness of tails. Downey (2000) suggests that the Log Normal (not heavy tailed in the classical sense), can fit duration distributions as well as classical heavy tail distributions, and gave some interesting physical motivation for this distribution as well. However, by developing the nice idea of "tail fragility", Gong, Liu, Misra and Towsley (2001) showed that both types of distribution can give apparently reasonable fits. This general direction was further developed by Hernández-Campos, Marron, Samorodnitsky and Smith (2002), using some much larger data sets (in the millions), together with a novel visualization for understanding the level of sample variation. This latter work showed that a mixture of three Double Pareto Log Normal distributions (see Reed 2001) gave an excellent fit, and was also physically interpretable. These results motivated the development of the concept of "variable heavy tails".

Another point of recent controversy has been over the issue of long range dependence. This is currently widely accepted (and intuitively sensible), but some interesting questions have been raised (using some novel visualization ideas) by Cao, Cleveland, Lin, and Sun (2001, 2002a, b, c). The key idea is that aggregated traffic, of the type typically found at major internet nodes, tends to "wash out" long range dependence. The idea is theoretically justified by appealing to limit theorems for aggregated point processes. An example, where both types of behavior were observed depending on scale, was studied using some different visualizations, by Hannig, Marron and Riedi (2001). An interesting issue to follow in the future will be the state of this balance between long range dependence caused by relatively few large extremely large transmissions, and a more Poisson type probability structure caused by aggregation. Cao, Cleveland, Lin, and Sun (2001, 2002a, b, c) predict ultimate Poisson type structure, for the good reason that internet traffic continually increases. However, this is based on an assumption that sizes of transmissions will stay fixed, which seems questionable.

Downey (2001) questioned long range dependence from a different viewpoint, by showing that duration distributions may not be very consistent with the definition of "heavy tailed", in the classical asymptotic sense. This was the first observation of "variable heavy tails", as defined in Hernández-Campos, Marron, Samorodnitsky and Smith (2002). That paper goes on to develop an asymptotic theory that parallels the classical theory. In particular it is seen that long range dependence still follows from the far broader (and realistic in terms of the nature of the data) concept of "variable heavy tails".

The main purpose of this paper is to point out some perhaps fun and challenging visualization problems.

The first main problem is related to the “Mice and Elephants” graphic, developed in Marron, Hernández-Campos and Smith (2002), discussed in Section 3. The problem is how to choose a “representative sample”, and it is seen that the usual device of random sampling is clearly inappropriate.

The second main problem is motivated by an apparent “commonality of flow rates”, discussed in Section 3. A large scatterplot seems to reveal some interesting visual structure, that makes physical sense. The question is how to best understand the driving phenomena.

2 Mice and Elephant plots and random sampling

The Mice and Elephants plot is a visualization that illustrates the fundamental theory discussed in Section 1. In particular, it shows how a heavy tailed distribution can lead to long range dependence, as explained below. This type of plot is shown in Figure 1. The key idea is that Internet “flows”, i.e. the set of packets that make up a single connection, are represented by line segments. The left (right, resp.) ends of the line segments show the times of the first (last, resp.) packets in each flow. Thus each line segment shows the “overall time of activity of that flow”. For good visual separation of the line segments, a random height is used on the vertical axis (essentially the “jitter” idea of Tukey and Tukey (1990), see also Cleveland (1993)).

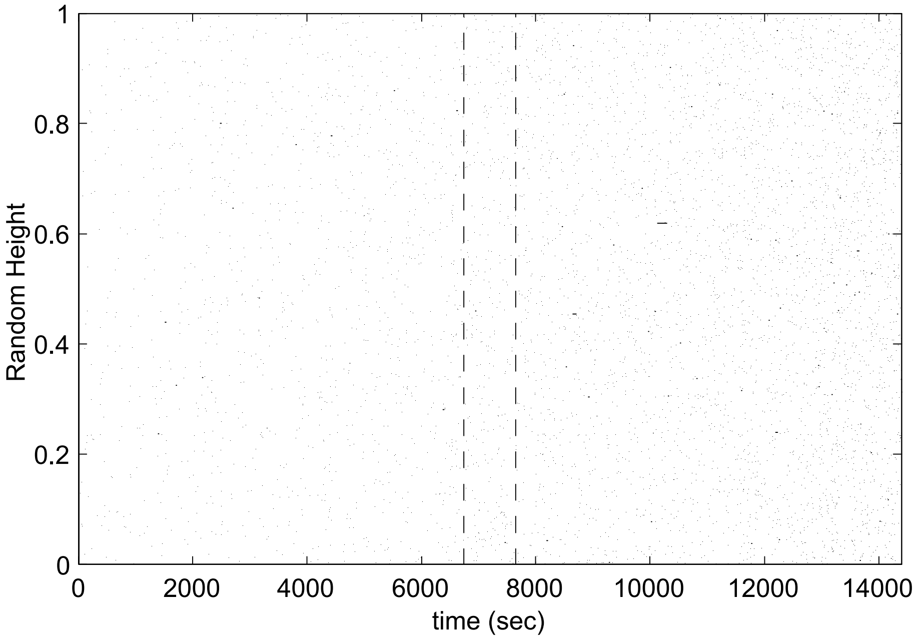

FIGURE 1: *Mice and Elephant plots for full four hour time block. Vertical bars indicate 15 minute time window shown in Figure 2. This suggests that all flows are "mice".*

The data here were HTTP (Web browsing) response times. They were collected during a four hour period, 8:00 AM - 12:00 noon, on a Sunday morning in April of 2001. This time period was chosen to represent a "light traffic" time. For a parallel analysis of a heavy traffic time, see Marron, J. S., Hernández-Campos, F. and Smith F. D. (2002). More detailed graphics for both analyses are available at the web address: http://www-dirt.cs.unc.edu/marron/MiceElephants/. For more details on the data collection and processing methods, see Smith, Hernández-Campos, Jeffay and Ott (2001).

The total number of flows for the time period in Figure 1 is 1,070,545. Massive overplotting resulted from an attempt to plot all of them. A simple and natural approach to the overplotting problem is to plot only a random subsample. This was done for a subsample size of 5000 (chosen for good visual effect) in Figure 1, and in the other figures in this section.

Figure 1 shows steadily increasing traffic, which is expected behavior on Sunday mornings (perhaps the times at which begin web browsing is driven by a wide range of adventures experienced on the previous night!). It also suggests that there are no long flows, with the longest visible flow being less than 5 minutes. This is a very serious mis-impression, that completely obscures the most important property of the traffic, as noted below

This point becomes clear from a similar graphic, but zoomed into the region between vertical bars in Figure 1, which represent the central 15 minutes (1/16th of the total time). Figure 2 shows this zoomed mice and elephants plot. There were 59,113 (not far from 1,070,545 / 16) flows that intersected this time range. Plotting all would again result in severe overplotting, so only a random sample of 5000 is plotted.

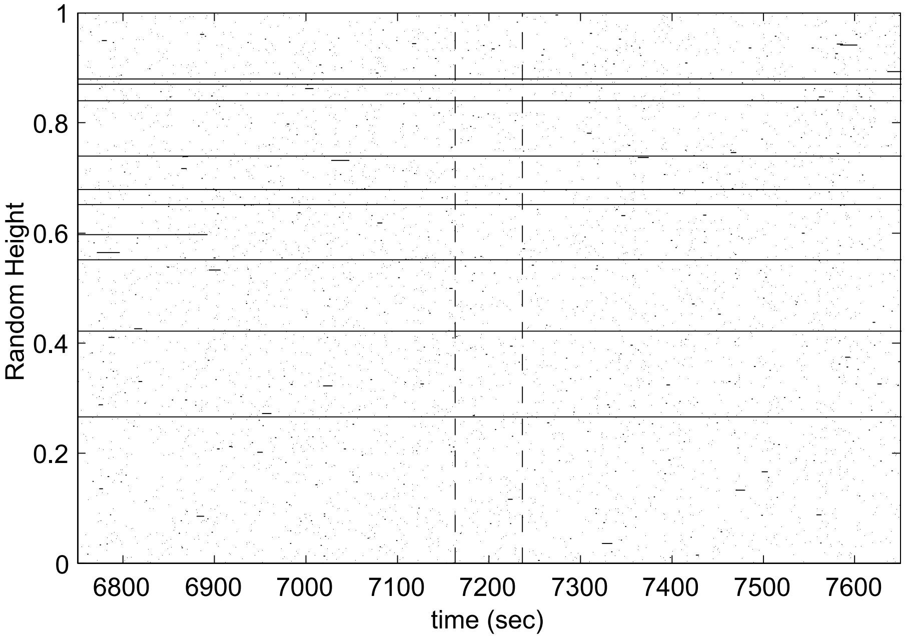

FIGURE 2: *Mice and Elephant plots for 15 minute time block. Vertical bars indicate time span containing 5000 flows, shown in Figure 3. This show both "mice" and "elephants".*

The visual impression of Figure 2 is far different from that of Figure 1. In particular, there are a number of flows that cross the full 15 minute interval, which seems quite contrary to the above impression that all flows are much less than 5 minutes in duration. This mis-impression is caused by a combination of the heavy tailed duration distribution and the random sampling process. Because of the heavy tails, there are only a very few flows that are very long. These have only a very small chance of appearing in the randomly selected sample. E.g. the chance that any of the largest 40 flows have a chance of appearing is only about $(40 \cdot 5000/1,070,545) \approx 0.05$. The number 40 is relevant, since 39 flows extend the full length of the central one hour time interval. This small probability of inclusion explains why none of these very long flows appear in Figure 1.

It is interesting to zoom in once again. Figure 3 shows the results of repeating the analysis for the region between the vertical bars in Figure 2. Those bars do not show 1/16th of the region in Figure 2, because that contains less than 5000 flows (which would give an inconsistent visual representation). Instead the bars are chosen so that exactly 5000 flows intersect the time interval (which is again centered in the range of the data), which is about 1.3 minutes long

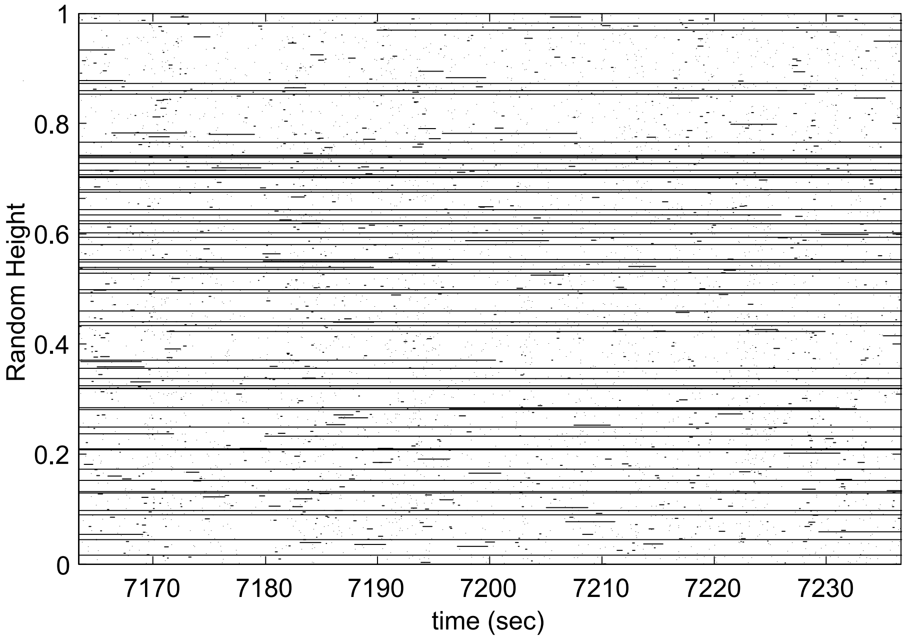

FIGURE 3: *Mice and Elephant plots for time window containing 5000 flows.*

Figure 3 shows a rather large number of long flows, and because there is no sampling, is perhaps "representative of behavior at a given time". However, this view is also biased because it shows in some sense "too high a proportion of long flows". The reason is a "length biasing" type of effect: long flows have a much greater chance of appearing in any such small interval, yet are a very small fraction of the population.

The clear conclusion from Figures 1, 2, and 3 stands in stark contrast to one of the most time honored principles of statistics (and a commonly used tool in visualization): simple random sampling of these data does not give a "representative sample". This problem is caused by the heavy tails, and there is a general principle at work: simple random sampling will never give a representative sample in heavy tail situations.

The first open problem proposed in this paper is to find an improved version of "representative sample". A sensible first step may be to decide what that means. Is there a reasonable mathematical definition of this that makes sense for heavy tailed distributions? Can classical length biased sampling ideas perhaps be useful? Are they visually interpretable?

Figures 2 and 3 show that the name "mice and elephants" is sensible for this graphic. It has become commonplace terminology in the internet research community for this phenomenon of a very few, very large flows. This concept is fundamental to the ideas outlined in Section 1. It is a clear consequence of the heavy tail duration distributions. It also makes the long range dependence in the aggregated time series visually clear. In particular, time series of binned traffic measures (such as packet counts) are essentially vertical sums of the line segments in the mice and elephants plots. The above described theory about heavy tails implying long range dependence is visually clear, given that the very long "elephant" flows clearly persist over quite long time ranges. It is not surprising that the persistence of the elephants results in the often observed "bursty behavior" of internet traffic.

Another interesting open problem is to use this visualization to motivate new quantitative measures for understanding the nature of this type of data. The standard notions of "heavy tails" for the duration distributions, and of "time series dependence", are not aimed at describing the full structure of this data. Instead they are just tools adapted from other areas, which perhaps result in a somewhat clumsy statistical analysis. Can the quantitative analysis be sharpened by quantitating other aspects of the full plot?

Mice and elephants visualizations also give a very clear view of the fact that the standard queueing theory models, with exponential duration distributions, are grossly inappropriate. This is seen in Figure 4, which is a duplicate of Figure 3, but for simulated data, with exponential distributions. To keep the comparison as fair as possible, the real data time range, sample sizes and even start times are used. Only the duration of each flow (the length of the line segment) is simulated. Also the exponential parameter is chosen to give a population mean that is the same as the sample mean for the real data.

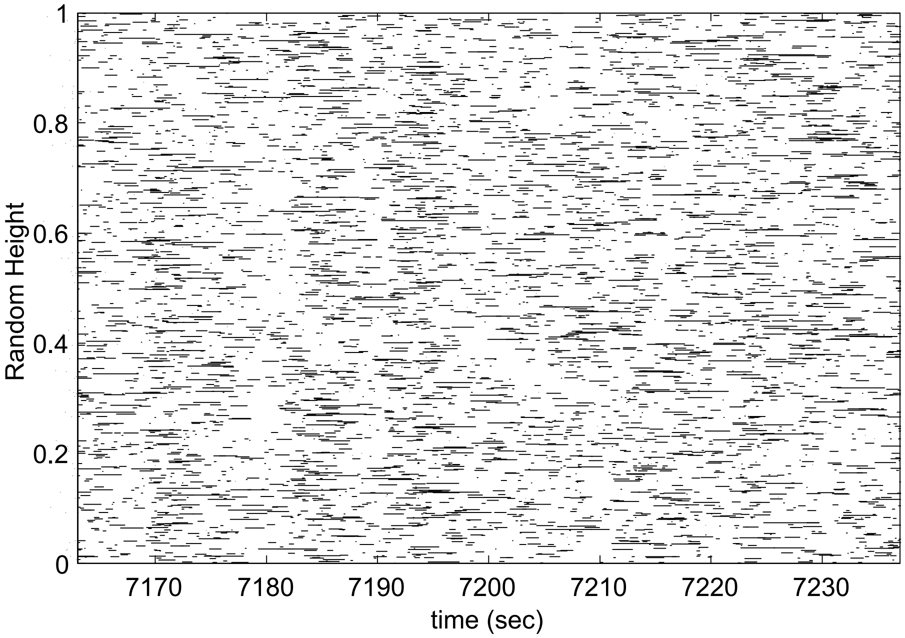

FIGURE 4: *Mice and Elephant plots for simulated exponentials, in setting of Figure 3.*

Figure 4 shows a completely different type of distribution of flow lengths, from the real data shown in Figure 3. In particular, there are no flows that are nearly long enough cover the whole interval, the number of very short flows is far fewer, and there are many more “medium size” flows. This is a consequence of the “light tail” property of the exponential distribution. Once the mean is specified, there are constraints on the frequencies of very large and very small observations. These constraints make the exponential distribution a very poor approximation to the type of behavior seen in Figures 1, 2 and 3. Thus these mice and elephants plots clearly illustrate the concept from Figure 1 that classical queueing models are inappropriate for internet traffic. In addition, the mice and elephants plot in Figure 4 seems quite consistent with the idea that when this traffic is vertically aggregated, the resulting time series exhibit only classical short range dependence.

The above proposed problem of how to subsample for effective mice and elephants visualization should not be regarded as “one off”. The reason is that the internet is constantly changing in many ways, and this could become a standard tool for monitoring change. In particular such monitoring could show the ultimate resolution to the above controversy, as to whether large scale aggregation will eventually swamp out long range dependence effects, or whether the latter will continue through the continued growth of elephants in frequency and size. An effective solution might also extend well beyond internet traffic, and become the foundations of a new theory of sampling in heavy tail contexts.

3 Commonality of flow rates

Another interesting view, of the HTTP response data analyzed in Section 2, is a scatterplot of the duration (time, i.e. length of the line segments) of each response, versus the size of the response in bytes (i.e. the amount of data transferred). Both variables share the heavy tailed “mice and elephants” behavior demonstrated in Figures 1-4, so a reasonable view of the data comes from plotting both variables on the log scale. Figure 5 shows the resulting scatterplot. This requires special handling of responses with 0 duration (e.g. this happens for single packet responses). This was done by dropping such responses from the sample, which resulted in 382,127 responses appearing in the scatterplot.

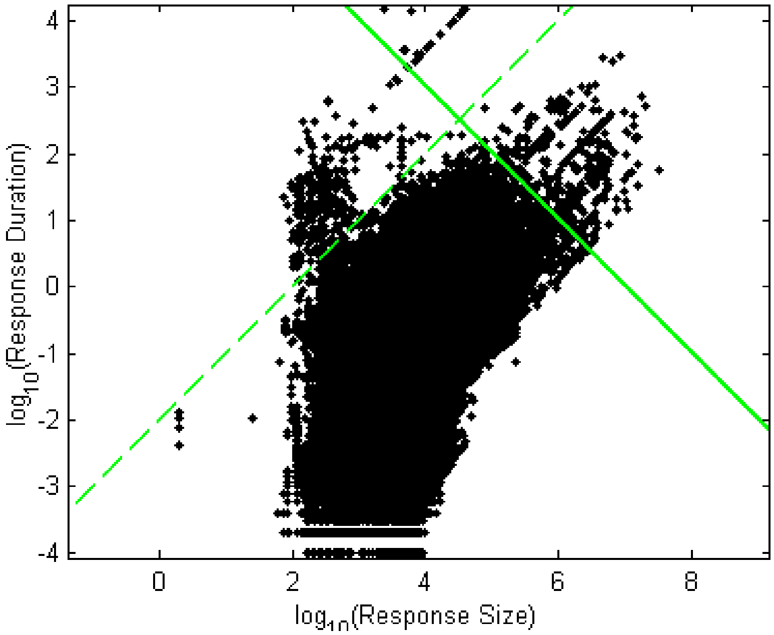

FIGURE 5: *Log - log scatterplot showing how flow duration time depends on transmission size. Suggests clusters of flows with same throughput (rates), as diagonal lines.*

The general tendency in Figure 4 is roughly what one might expect: larger size responses need more time, so there is a general upwards trend. Horizontal lines at the bottom of the plot reflect discreteness of very small time measurements. A perhaps surprising feature is the diagonal lines of points present at larger times and sizes. Not only do the lines appear to be parallel, they also lie at a 45° angle to the coordinate axes, as indicated by the parallel dashed green line, which has equation $y = x - 2$. These diagonal lines of points represent sets of flows where

$$\log_{10} time = \log_{10} size + C,$$

for some constant C, which is the same as

$$size = R \cdot time,$$

where $R = 10^{-C}$ is interpretable as a "constant rate". Thus the flows following each diagonal line have essentially the same rate (defined as total size divided by total rate).

Figure 5 gives a strong visual impression that the large flows may be "naturally clustered" in terms of rates. This is sensible because rates are naturally driven by the nature of the network between the source and the destination. Most of the computers within UNC will likely have quite similar rates to a few popular web-sites, resulting in similar rates for large numbers of transfers.

The second main open problem of this paper is to develop methods for analyzing this aspect of the population. How can the clusters be isolated? What are the cluster boundaries? How many flows are in the major clusters?

A start on addressing these issues appears in Figure 6. Here the data are projected onto the orthogonal solid green line in Figure 5, so the problem is reduced to studying clusters in univariate data. For easy visual connection to Figure 5, the data are transformed to the coordinate system which allows treating the solid green line as the axis. In particular the transformation is:

$$proj = -2 - \frac{\log_{10} time - \log_{10} size}{\sqrt{2}}.$$

The denominator of $\sqrt{2}$ makes the transformation length invariant (i.e. a rotation), and the subtraction from -2 gives the most straightforward view of the solid green line as an axis.

The top panel of Figure 6 shows two displays of the projected data. The first is the green dots, which are a standard “jitter plot” (again see Tukey and Tukey (1990) and Cleveland (1993)), where the horizontal coordinate is the projection value (i.e. location of each data point when projected onto the green line), and a random vertical coordinate is used for visual separation. The jitter plot shows only a random sample of 10,000 to avoid overplotting problems. The second display of the data is the family of blue curves. These are kernel density estimates (essentially smooth histograms), with a wide range of window widths. Looking at a family of smooths is the “scale space view” of data, which is recommended as a practical solution to the traditional problem of bandwidth choice, see Chaudhuri and Marron (1999) for further discussion.

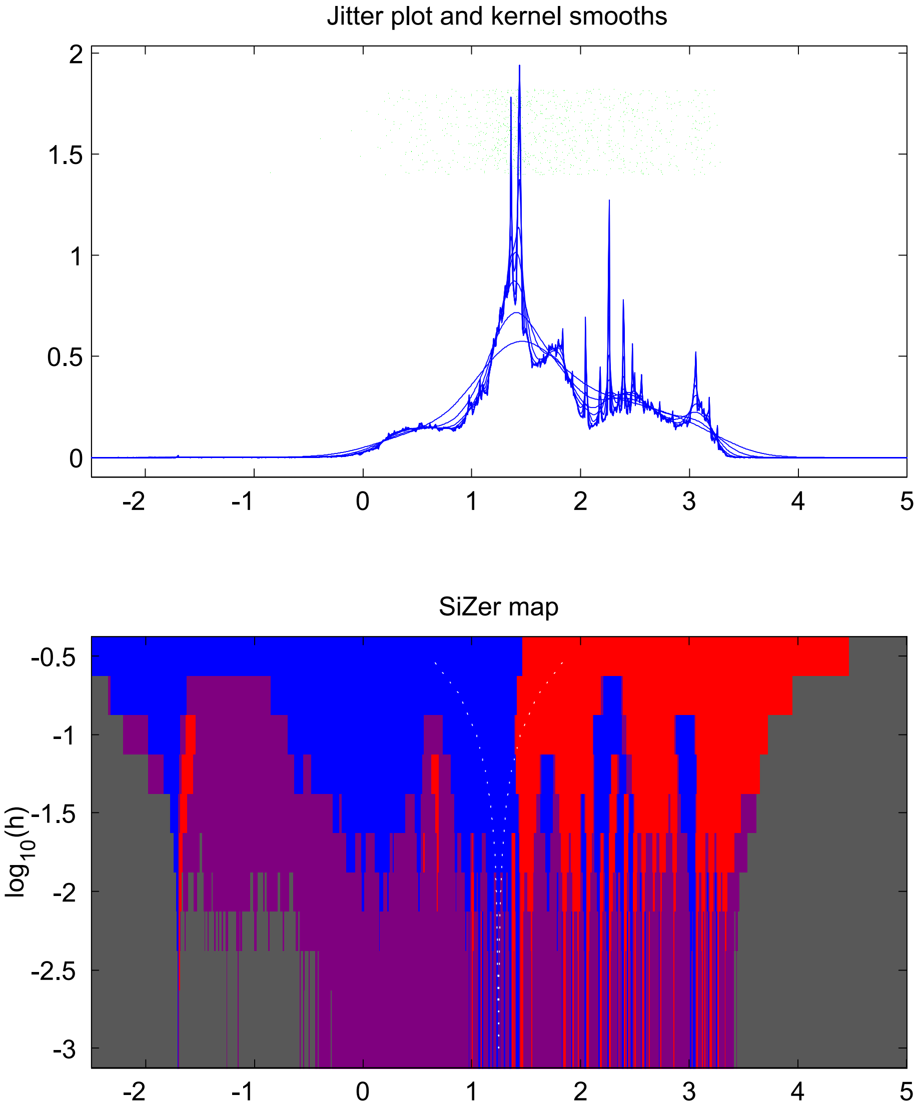

FIGURE 6: *SiZer analysis of projected scatterplot. Show many significant clusters.*

The family of kernel smooths suggests a number of "broad bumps", and there are also a number of "small spikes". It is tempting to dismiss the spikes as "spurious sampling variability", however recall that such "clusters" were suggested in Figure 5, and a possible physical explanation was suggested above. Furthermore the sample size $n = 382,127$ is fairly large, so perhaps those spikes represent "important underlying structure" in the data?

A useful tool for addressing such exploratory data analysis questions is the SiZer map shown in the bottom panel of Figure 6. Rows of this map correspond to different window widths, i.e. to blue kernel smooths, and the horizontal axis is the same as in the top panel. Colors are used to indicate statistical significance of the slopes of the blue curves, with blue (red, resp.) for significantly increasing (decreasing, resp.), with purple for regions where the slope is not significantly different from 0, and with gray where the data are too sparse for reliable inference.

The SiZer map shows that all of the “broad bumps” are statistically significant, as are most of the tall thin bumps. These may not be surprising because $n = 382,127$ allows resolution of quite a few features of the underlying probability density, in view of the large sample size. More surprising may be the very small bump at -1.7. This is hardly visible in the blue family, and yet is clearly statistically significant in the SiZer map.

The analysis of Figure 6 is not very satisfying, because it seems that perhaps some of the clusters, that are clearly visible as lines of points in Figure 5, might be “masked” by the large amount of other data that makes up the “broad peaks”. A simple approach to this is to repeat the analysis for a suitably threhsolded sub-sample. Visual inspection of Figure 5 suggests using only the data above the solid green line, $y = -x + 7$. There were 572 such points, still enough for effective kernel density estimation. The resulting analysis is shown in Figure 7.

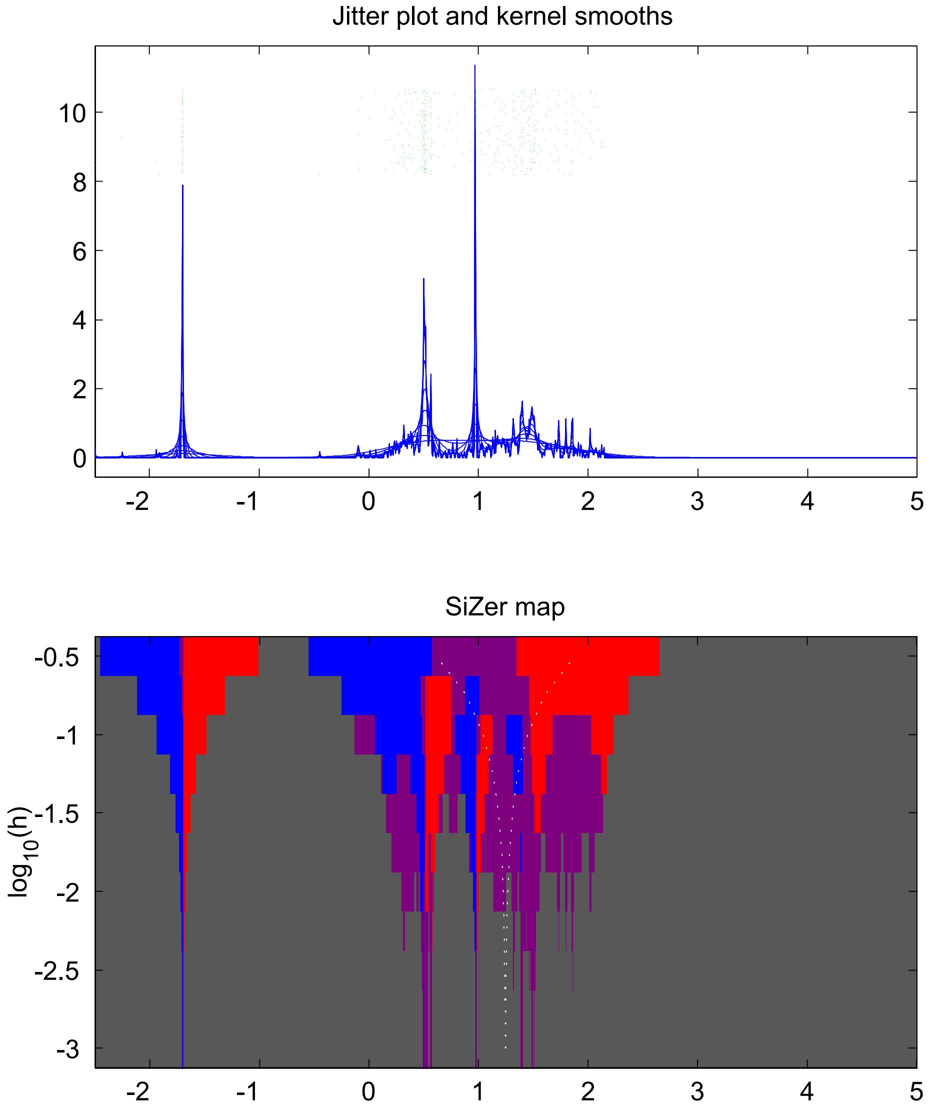

FIGURE 7: *SiZer analysis of scatterplot points above the solid green line in Figure 5. Shows significant clusters, but different from those of Fiugre 6. Thus many more clusters than those in Figure 5 exist.*

As expected, the "broad bumps" in Figure 6 have now disappeared. There are also some very significant "slim spikes". Note that the spike near -1.7 is now much taller in the blue family of curves (essentially all of these data points have been retained from Figure 6, and now are proportionally a far larger part of the population). However, note that many of the tall thin peaks in Figure 6 are not present in Figure 7. This shows that much of the "clustered aspect" of the population actually occurs more in the main body of the main scatterplot in Figure 5, and thus can not be teased out by simple thresholding as done in Figure 7.

Thus, this is a case where the scatterplot of Figure 5 hides a large amount of interesting population structure. The SiZer analysis is an indirect way of understanding this. Are there more direct ways of visualizing this type of structure?

4 Acknowledgements

The author is grateful to the Cornell University School of Operations Research and Industrial Engineering for hospitality, support, and an exciting research environment. This research was supported by Cornell University's College of Engineering Mary Upson Fund and NSF Grant DMS-9971649.

References

[1] Cao, J., Cleveland, W. S., Lin, D. , and Sun, D. X. (2001) On the Nonstationarity of Internet Traffic, *Proceedings of the ACM SIGMETRICS '01*, 102-112. Internet available at: http://cm.bell-labs.com/cm/ms/departments/sia/wsc/webpapers.html.

[2] Cao, J., Cleveland, W. S., Lin, D. , and Sun, D. X. (2001) The Effect of Statistical Multiplexing on the Long Range Dependence of Internet Packet Traffic, Bell Labs Tech Report, 2002, Internet available at: http://cm.bell-labs.com/cm/ms/departments/sia/wsc/webpapers.html.

[3] Cao, J., Cleveland, W. S., Lin, D. , and Sun, D. X. (2001) nternet Traffic: Statistical Multiplexing Gains, *DIMACS Workshop on Internet and WWW Measurement, Mapping and Modeling*, Internet available at: http://cm.bell-labs.com/cm/ms/departments/sia/wsc/webpapers.html.

[4] Cao, J., Cleveland, W. S., Lin, D. , and Sun, D. X. (2001) Internet Traffic Tends Toward Poisson and Independent as the Load Increases, *Nonlinear Estimation and Classification*, eds. C. Holmes, D. Denison, M. Hansen, B. Yu, and B. Mallick, Springer, New York, Internet available at: http://cm.bell-labs.com/cm/ms/departments/sia/wsc/webpapers.html.

[5] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807-823.

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

[7] 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.

[8] Crovella, M. E. and A. 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.

[9] Downey, A. B. (2000) The structural cause of file size distributions, Wellesley College Tech. Report CSD-TR25-2000. Internet available at: http://rocky.wellesley.edu/downey/filesize/.

[10] 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/.

[11] 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.

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

[13] 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.

[14] Hernandez-Campos, F., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2002) Variable Heavy Tailed Durations in Internet Traffic, unpublished manuscript, web available at http://www-dirt.cs.unc.edu/marron/VarHeavyTails/.

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

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

[17] Marron, J. S., Hernández-Campos, F. and Smith F. D. (2002) Mice and Elephants Visualization of Internet Traffic, submitted to proceedings of the CompStat 2002, internet available at: http://www-dirt.cs.unc.edu/marron/MiceElephants/.

[18] 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/.

[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, 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.

[22] Tukey, J., and Tukey, P. (1990). Strips Displaying Empirical Distributions: Textured Dot Strips. Bellcore Technical Memorandum.