Mice and Elephants Visualization of Internet Traffic

J. S. Marron[1], Felix Hernandez-Campos[2] and F. D. Smith[2]

[1] School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 14853, USA and Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

[2] Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599-3175, USA

Abstract. Internet traffic is composed of flows, sets of packets being transferred from one computer to another. Some visualizations for understanding the set of flows at a busy internet link are developed. These show graphically that the set of flows is dominated by a relatively few "elephants", and a very large number of "mice". It also becomes clear that "representative sampling" from heavy tail distributions is a challenging task.

Keywords. Heavy Tail Distributions, HTTP Flows, Visual Representation, Zooming Graphics.

1 Introduction

Internet traffic is composed of a large number of small "packets" flowing from one computer to another. These packets are organized into flows that represent blocks of data, such as file transfers. In this paper, a visualization is developed for studying the nature of the population of flows passing through a given internet link.

An example of the visualization is shown in Figure 1. The behavior of flows over time is shown by representing each flow as a horizontal line segment, whose left end point is the start time of the flow (i.e. the time of the first packet), and whose right endpoint is the end time of the flow (the time of the last packet). Thus each line segment represents the full time period where the flow is "on". For convenient visual separation of a reasonable number of flows, a random height is added. This follows the spirit of the jitter plot idea of Tukey and Tukey (1990). See also pages 121-122 of Cleveland (1993).

An interesting feature of the flows shown in Figure 1 is a very large number of very short flows, and a few very long flows. The terminology "mice and elephants" provides a useful metaphor for understanding this characteristic of internet traffic: there are relatively few elephants, and a large number of mice. Another way of thinking of mice and elephants is that the distribution of lengths of the flows is heavy tailed.

The data shown in Figure 1 are HTTP (Web browsing) response times. They were collected during a four hour period on a Thursday afternoon in April of 2001. This time period was chosen to represent a "heavy traffic" time. An HTTP response time is defined here as the time between the first and last packets of a single HTTP data transfer. For more details on the data collection and processing methods, see Smith, Hernandez-Campos, Jeffay and Ott (2001).

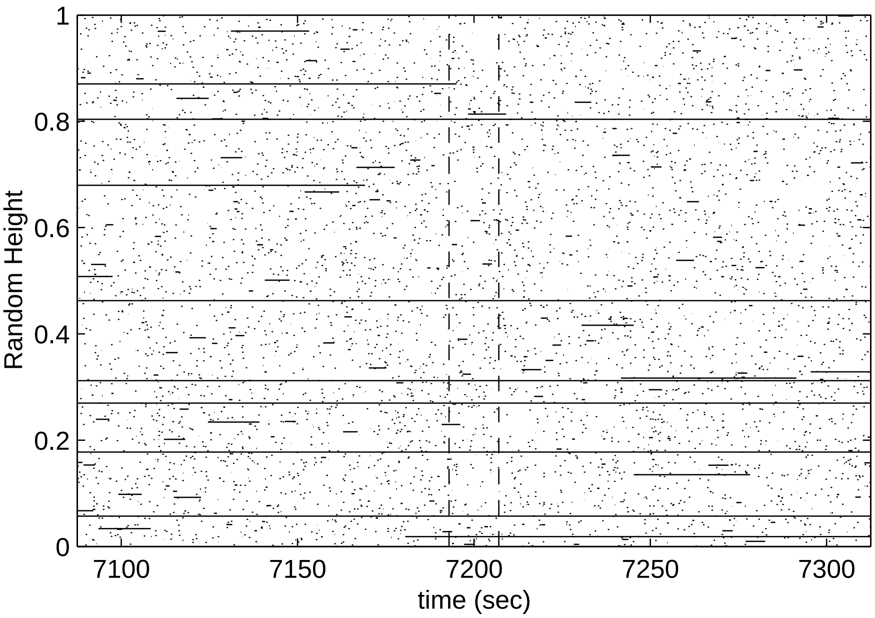

Fig. 1. Mice and Elephant display, showing HTTP flows. This is a random sample of 5000 out of 104,839 flows overlapping a 3.75 minute time interval.

During the four hour time window, there were 6,870,022 total HTTP responses. If all of these were shown in Figure 1, then the image would be useless, because it would be totally black. Sub-sampling is a natural approach to this problem.

One way to sub-sample is to reduce the time window. This is done in Figure 1, where only flows which spend at least part of their life times in the shown interval are considered. The interval has length 3.75 minutes = 4 hours / 4^3 (the reason for this choice will become apparent in Section 2), and is centered in the full 4 hour range of the data. The vertical dashed bars indicate the window of a further time restriction, that will be discussed in Section 2. When the full 6,870,022 responses were sub-sampled to the window of Figure 1, there were still *104,839* left. This is still far too many to show all of the mice and elephants, without heavy overplotting.

Another approach to reducing the number that are shown is random sampling. This was also done in Figure 1, where only a random sample of 5000 is shown. The number 5000 was the result of some experimentation with the goal of pleasing visual effect. Larger numbers made some of these plots “too busy”, and a smaller number seemed inefficient in terms of not showing enough of the data.

Random sampling has very broad appeal as a means of finding a “representative subsample”. Indeed this notion lies at the heart of a large number of statistical procedures in routine use. However, random sampling has some very serious drawbacks in the present case, because the distribution of lengths is heavy tailed. In particular, since there are relatively quite few elephants, their chance of appearing in the random sample can be quite negligible, which means they may not be properly represented in the visualization. This problem is investigated in Section 2. The problem becomes very clear from looking across a range of different "scales" (time windows). An alternative visualization is discussed in Section 2.

This mice and elephants visualization also provides visual insight into the causes of the interesting and compelling theory that has been built up to explain phenomena that have been observed in internet traffic. This is discussed in Section 3.

2 Deeper Look at Sampling Issues

In this section, sampling issues are carefully investigated. The key insights come from the construction of mice and elephant plots over a wide range of different time scales. An interesting sequence of views of this type may be found following the link "Zoomed-in Views of Mice & Elephants in Thursday Afternoon Trace" from the web page

`http://www-dirt.cs.unc.edu/marron/MiceElephants/`.

To save space, only a carefully selected few are shown as figures in this paper.

Figure 2 show the mice and elephants view over the full four hour time interval. This can be thought of as "zooming out in time scale". Once again only a random sample of 5000 is displayed (to avoid massive overplotting). The parallel bars near the center show the 3.75 minute time interval that is used in Figure 1 (thus giving a visual impression of the amount of "zooming out" that has been done). Figure 2 gives the impression that the traffic is all mice, and there are no elephants, which is clearly different from the impression of Figure 1. In particular, Figure 1 shows several elephants which span the full time interval, while none of these are visible over the same time interval in Figure 2.

The differences between Figures 1 and 2 are explained by the sampling process. In particular, as noted by a number of authors (referenced in Section 3), this distribution is quite "heavy tailed", which means that one may expect some very large values, but there will be only a very few of them. In particular, a careful look showed that there were 116 flows (in the full data set), that covered the full time range in Figure 1. Several of these show up in Figure 1, because there 5000 flows were randomly selected from the 104,839 that intersected that time interval, i.e. the subsample was about 5% of the total. Hence it is not surprising that about 6 of the 116 flows that cross the full interval appear in Figure 1. But the sampling situation is far different in Figure 1, where the 5000 flows were drawn from the full population of *6,870,022*, i.e. the subsample is less than 0.1% of the total. Hence it is not surprising that none of the 116 flows that completely cross the time interval of Figure 1, appear in Figure 2.

Figure 3 show the effect of "zooming in" (from the view shown in Figure 1) with respect to the time scale. Here the time interval is reduced to only the rather short 14 second interval between the vertical dashed bars in Figure 1. Once again, the view is radically different from that shown in Figure 1. This time the visual impression is of far more elephants. This view even conveys the impression that mice are a fairly negligible component of the population!

As above, this mistaken impression of the full population is the result of random sampling. This time 5000 flows were randomly selected from the 6910 that intersect this time interval, i.e. about 83%. This time interval is

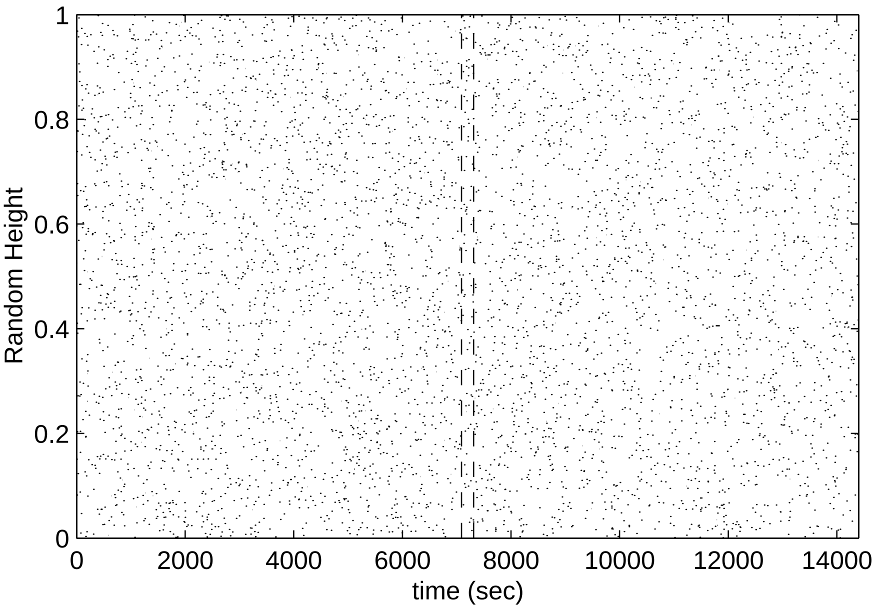

Fig. 2. Mice and Elephant display, over a long 4 hour time interval. This is a random sample of 5000 out of 6,870,022 total flows.

completely crossed by 163 flows (this might be viewed as surprisingly few, since 116 flows crossed the 16 times longer interval shown in Figure 1). Since most of these appear in the sample shown, the visual impression is dominated by these elephants. Indeed a reasonable opinion is that overplotting has occurred in Figure 3.

Figure 4 shows a sense in which Figure 3 is rather representative of the data. In Figure the full 4 hour time interval is shown, but instead of random sampling, the longest 5000 flows are shown. This is called the corresponding "elephant plot". This provides another way of seeing that this distribution of flow lengths is very heavy tailed. As noted above, this is less than 0.1% of the total data, and yet the smallest flows shown appear as dots (no visible length at this time scale). This means that the other 99.9% of the data would also appear as dots (if they could all be usefully plotted).

It is also interesting to view modifications of Figure 4, which zoom through smaller values of the threshold k, where the largest k flows are shown (thus $k = 5000$ is shown in Figure 4). This can be viewed following the link "Elephants-only Views for Thursday Afternoon Trace" from the above web page.

The Thursday afternoon time period shown here was chosen to represent a peak traffic time. Similar plots were also constructed for the off peak time period of Sunday morning.. The mice and elephants plots, as in Figures 1, 2 and 3, can be found following the link "Zoomed-in Views of Mice & Elephants in Sunday Morning Trace", and some elephant plots, as in Figure 4, can be found following the link "Elephants-only Views for Sunday Morning Trace"

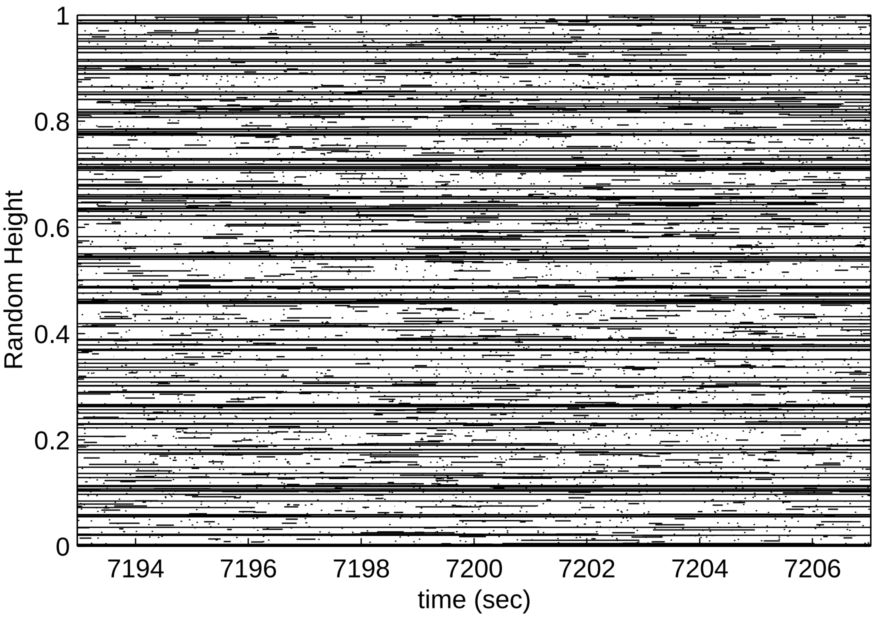

Fig. 3. Mice and Elephant display, showing HTTP flows over a very narrow time window. This is a random sample of 5000 out of 6910 flows overlapping a 14 second time intervals.

from the above web page.

A clear lesson is that random sampling is not an effective method for choosing a "representative sample" for internet data, or indeed data from heavy tail distributions that appear elsewhere, e.g. financial and environmental contexts. This is an interesting exception to the usual lesson taught in elementary statistics courses, in the special case of heavy tail distributions. An interesting open problem is how can a representative sample be chosen, for this type of visual impression, in heavy tailed contexts? There are related open sub-sampling problems (with perhaps different answers), for other purposes. For example, internet traffic data sets in generally tend to be large, and often need to be subsampled to efficiently store for later analysis and comparison. How can such a sample be made "representative"?

3 Related Theory

The mice and elephants plots presented in this paper provide visual insight into some important theoretical results often applied to Internet traffic data. The main idea is that heavy tailed flow length distributions give rise to long range dependence in the aggregated traffic.

The heavy tail of the flow length distribution has been observed by Garrett and Willinger (1994), Paxson (1994), Crovella and Bestavros (1996). See Downey (2001) for an alternative view, and some apparently contradictory ideas. The controversy raised by Downey was resolved by Gong, Liu, Misra

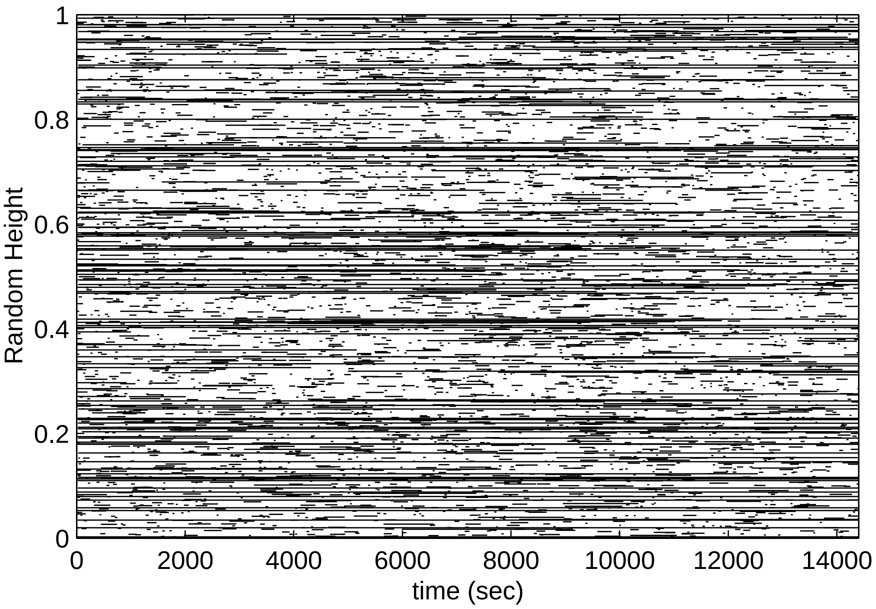

Fig. 4. Elephants display, showing the 5000 largest HTTP flows, over the longer 4 hour time window.

and Towsley (2001) and by Hernandez-Campos, Marron, Samorodnitsky and Smith (2002). The latter paper is a particularly detailed study, with more than usual attention paid to statistical issues, such as careful accounting for sampling variability.

Beautiful mathematics showing how heavy tailed flow lengths lead to long range dependence can be found in 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), Resnick and Samorodnitsky (1999). A recent and general version of this can be found in Hernandez-Campos, Marron, Samorodnitsky and Smith (2002).

Plots such as Figure 1 give a clear intuitive view of how a few elephants create long range dependence. The corresponding short range situation is illustrated in Figure 5. Figure 5 is based upon a simulated set of flows, whose starting times are the same as those in the original full data set ($n = 6,870,022$). However, the lengths of the flows are simulated from an exponential distribution, where the mean is taken to be the sample mean from the mean lengths. The exponential flow length distribution is the foundation of classical queueing theory, and is well known to result in short range dependent aggregated traffic. Although the mean flow length is the same as in Figure 1, the visual impression of the flows in Figure 5 is radically different: there are no elephants. How can the mean lengths be the same when there are no elephants? This happens because the mice in Figure 1 are smaller than those in Figure 5. This is not visible here, but can be seen by zooming in further, follow the link "Zoomed-in Views of Mice & Elephants in

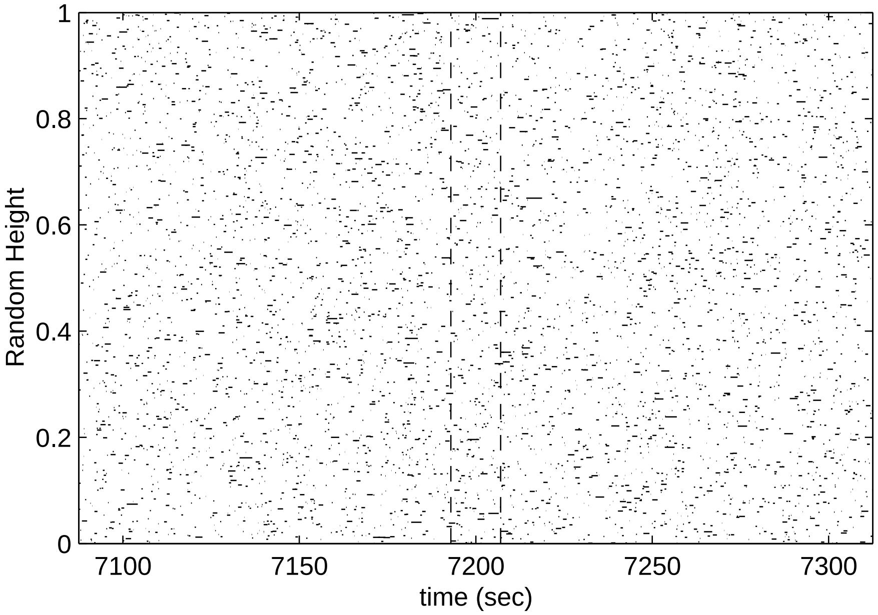

Fig. 5. Mice and Elephant display, showing simulated Exponentail (with the same mean) HTTP flows over the same time window as Figure 1.

Synthetic Version of Thursday Afternoon Trace" from the above web page.

Because there are no elephants in Figure 5, when the traffic is aggregated into bins, the correlations between the counts fall off quickly. However, when a few elephants, such as shown in Figure 1 enter the picture, the correlations between bin counts die off far more slowly, creating long range dependence.

An aspect of the distribution of flows not studied here is the distribution of the flow starting times. Recent analysis of these may be found in Cleveland, Lin and Sun (2000), Nuzman, Saniee, Sweldens and Weiss (2002) and Marron, Hernandez-Campos and Smith (2001).

References

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

Cleveland, W. S., Lin, D. and Sun, D. X. (2000) IP packet generation: statistical models for TCP start times based on connection-rate superposition, *Performance Evaluation Review: Proc. ACM Sigmetrics 2000*, 28, 166-177.

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.

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.

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/`.

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

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`.

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 `http://www.stat.unc.edu/postscript/papers/marron/NetworkData/LogNorm2LRD/`

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.

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.stat.unc.edu/postscript/papers/marron/NetworkData/VarHeavyTails/`.

Marron, J. S., Hernandez-Campos, F. and Smith, F. D. (2001) A SiZer analysis of IP Flow start times, unpublished manuscript.

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.

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

Nuzman, C., Saniee, I., Sweldens, W. and Weiss, A. (2002) A compound model for TCP connection arrivals for LAN and WAN applications, unpublished manuscript.

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

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.

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

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.

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