Journal of the Korean Statistical Society (2001), **30**: *2, pp 327–345*

Zooming Statistics: Inference across scales

Jan Hannig [1], **J. S. Marron** [2] **and R. H. Riedi** [3]

ABSTRACT

New statistical methods are needed to analyze data in a multi-scale way. Some multi-scale extensions of standard methods, including novel visualization using dynamic graphics are proposed. These tools are used to explore non-standard structure in internet traffic data.

1. Introduction

A topic of keen current interest in many areas of mathematical modelling, is integrating models across a broad range of scales. There are serious challenges in this area, because reasonable models exist for most single scale phenomena, however they often do not interface well with each other across wide ranges of scale. In particular, critical and useful approximations made at one scale frequently break down completely at other scales.

The major research effort underway in multiscale mathematical modelling spawns the need for statistical research in very wide scale settings. There is a strong need for both exploratory and confirmatory multiscale methodologies. This paper proposes some wide scale extensions of the standard time series method of autocorrelation analysis, and of the newer exploratory data analysis method of SiZer.

Dynamic graphic visualization is the key to the methods proposed in this paper. Internet traffic data analysis provides an excellent testbed for the illustration of this methodology, because large amounts of data are available, and there is interest in the structure of the data across a wide range of scales. The data set used in this paper is described in Section 2. In Section 3, the zooming autocorrelation method is developed, and used to highlight some non-standard (from the viewpoint currently accepted in that area) structure in the internet

[1]Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA

[2]Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

[3]Department of Electrical and Computer Engineering, Rice University, MS 380, Houston, TX 77251-1892

traffic data. This unexpected structure is explained by some simple covariance calculations. A complementary analysis, illustrating the zooming SiZer method, is done in Section 4.

A much different approach to multiscale time series analysis is using wavelets. See Percival and Walden (2000) for an introduction to that literature.

2. Internet Traffic Data

A major challenge for engineers, for computer scientists, for statisticians and for probabilists is the analysis and modelling of internet traffic data. The problem is of central importance because the present protocols were not designed with today's massive scale of the world wide web in mind, which results in large inefficiencies. A major research effort is under way to find improvements. For both the developmental and the confirmatory phases of this work, traffic models and analysis tools are vital. Really new ideas and models are needed because heavy tailed distributions and long range dependence (both appearing at a number of different points) render standard methods, such as classical queueing theory, unusable.

Interesting and important behavior has been observed at a wide variety of points on the internet. The first data set analyzed in this paper, consists of traffic measured on the main internet link to the University of North Carolina in 1998. Similar data sets are available online at the National Laboratory for Applied Network Research, at the web address http://moat.nlanr.net/PMA/. The data considered here are a series of time stamps, representing the arrival times of individual packets, for one million packets. In 1998, the traffic on this link was such that it took about 3 minutes to gather this data set.

The first models considered for internet traffic data were the standard queueing theory models, based on independent Poisson arrival times, and exponential waiting times. It was natural to try these first, as they were very successful for modelling e.g. telephone call traffic. But in a landmark paper, Paxson and Floyd (1995) made it clear that such models are inappropriate for internet traffic. This started a wide ranging search for more appropriate models.

An important difference between observed internet traffic and predictions from the Poisson process is the presence of long range dependence in the data, which is studied carefully in this paper. Cao, Cleveland, Lin, and Sun (2001) made the interesting observation that at small time scales, on a main internet link, the packet interarrival process is well approximated by independent expo-

nential random variables (as for a Poisson process). They correctly argue that for studying queueing behavior, small time scale behavior is the critical driving factor.

At first glance, this latter work seems to be in contradiction with the above observed long range dependence. Exact Poisson processes maintain their independent increment structure, regardless of the scale of aggregation. But Cao, Cleveland, Lin, and Sun (2001) point out that they have only a small scale model, and acknowledge the presence of long range dependence at larger time scales. Many of the long range dependent models being mentioned above, e.g. those in Feldman, Gilbert and Willinger (1998), Riedi, et. al. (1999) and Riedi and Willinger (1999), allow the accommodation of scaling behavior which changes with scale, but they do not explore the actual auto-correlation over a range of scales.

Deeper understanding of internet data can come from a widely cross scale analysis which includes both of these time scales. This is done in a simple way for the UNC arrival process data, by analyzing bin counts, for a wide range of binwidths. In this paper, linear binning, as described in Fan and Marron (1994) is used at all points. The binwidth m (measured in seconds) will parametrize the scale under consideration.

The autocorrelation structure of the time series of bin counts, is considered in a multiscale way in Section 2. The multiscale behavior of the autocorrelation function is quite different from what could be expected assuming that particularly strong dependence was present around round trip time. Thus the autocorrelation should be large at lags corresponding to round trip time. When the scale is increased, these lags move from right to left, so it is expected that a "lump of large correlation" will move from right to left. Instead a far different multiscale behavior is observed. In particular, there is a "constant lifting" of the autocorrelation across lags, showing that the concept of "long range dependence occurring at certain scales" is fallacious, and needs reconsideration. This apparent contradiction is explained by some heuristic calculations.

For some large scales, it is seen that sampling variability is an issue. Hence, a second larger data set, gathered over a longer time scale of nearly one hour will be considered. Here the data are separately binned over a range of time scales that increase by a factor of 2 each time. Each successively larger scale uses the previous data, plus an equal amount of new data.

SiZer provides a different type of exploratory data analysis, based on smoothing the data (using a family of scatterplot smoothers on the bin counts) to study local trends. This analysis is done in Section 4, where a similar cross scale

behavior is observed.

3. Zooming Autocovariance

Given a time series, $\{X_t : t = 1, ..., n\}$, the autocorrelation function, at lag $l = 0, ..., n-1$ is

$$\rho(l) = \frac{cov\left(X_\bullet, X_{\bullet - l}\right)}{var\left(X_\bullet\right)},$$

where

$$cov\left(X_\bullet, X_{\bullet - l}\right) = \frac{1}{n}\sum_{t=l}^{n}\left(X_t - \overline{X}\right)\left(X_{t-l} - \overline{X}\right),$$

$$var\left(X_\bullet\right) = cov\left(X_\bullet, X_{\bullet - 0}\right),$$

and

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t.$$

As noted in any time series text, e.g. Brockwell and Davis (1991), the autocorrelation function is a standard method for quantifying dependence in the statistical analysis of time series. In the case of mean zero Gaussian processes, the autocorrelation function $\rho(l)$ completely determines the distribution of the time series.

The best way to study how the autocorrelation function, of the bin counts time series, changes across scale is to view a movie (i.e. a dynamic graphical representation). Such a movie is internet available as the file ZoomStatFig1.mpg in the web directory

`http://www.unc.edu/depts/statistics/postscript/papers/marron/ZoomStat/`

If at all possible, it is recommended that this movie be viewed now. For discussion here, some frames of this movie are shown as Figures 1a-d.

Figure 1a shows the smallest scale binwidth, $m = 10^{-3}$ sec considered here. The blue horizontal line is the x-axis, included because it highlights the fact that the autocorrelation is always positive for the range of lags l that are shown. The autocorrelation is "generally quite low", which fits with the observations of Cao, Cleveland, Lin, and Sun (2001). However, the red dashed line, which is the 95th percentile of the autocorrelation for a Gaussian white noise process, shows that the autocorrelation is significantly different from that of a truly independent process.

Figure 1b shows the same analysis (the colored lines have the same meaning as in Figure 1a) for the larger scale of $m = 10^{-2}$ sec. The blue bar is included to graphically represent the scale (especially important in the movie version). This is now in the time scale neighborhood of one TCP round trip time (the scale at which long range dependence has been reported), so it is expected that long range behavior is encountered. But a careful look at the movie reveals a surprising feature: the long range dependence does

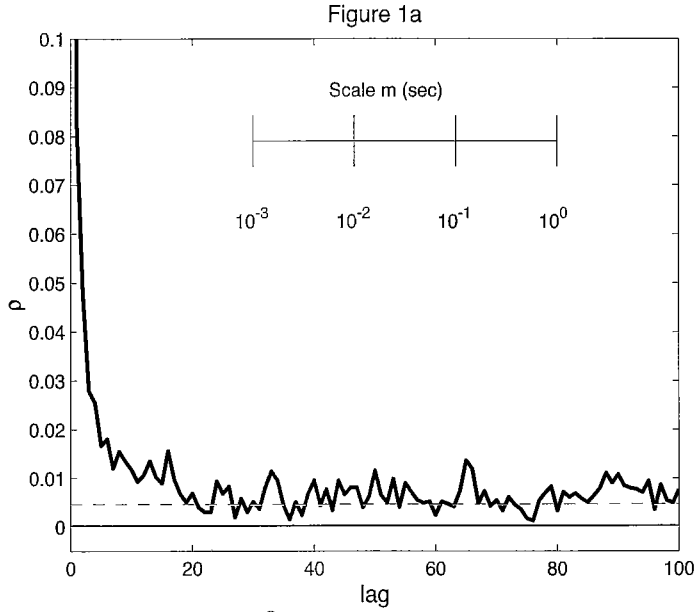

Figure 1a: Smallest scale, $m = 10^{-3}$, autocorrelation plot of 3 minute Internet traffic data set. Shows bin counts not far from independent.

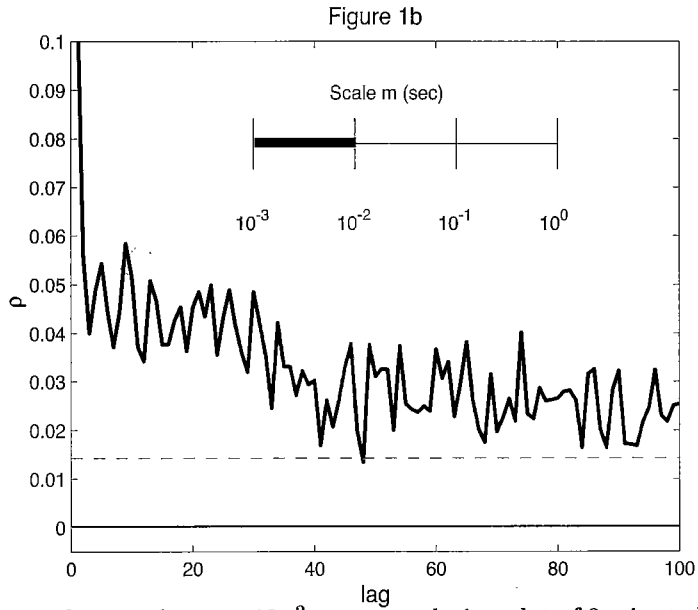

Figure 1b: Intermediate scale, $m = 10^{-2}$, autocorrelation plot of 3 minute Internet traffic data set. Shows bin counts with large positive autocorrelation.

not "move in from the right" as expected. This is expected because the time span corresponding to lag l, at scale m, is $l \times m$ (sec). Thus if "dependence occurs at time scale m_0", then as m is increased, the "hump of dependence", should move into the

picture from the right. However instead the autocorrelation function just "lifts vertically in a level fashion" to the point shown in Figure 1b.

Figure 1c

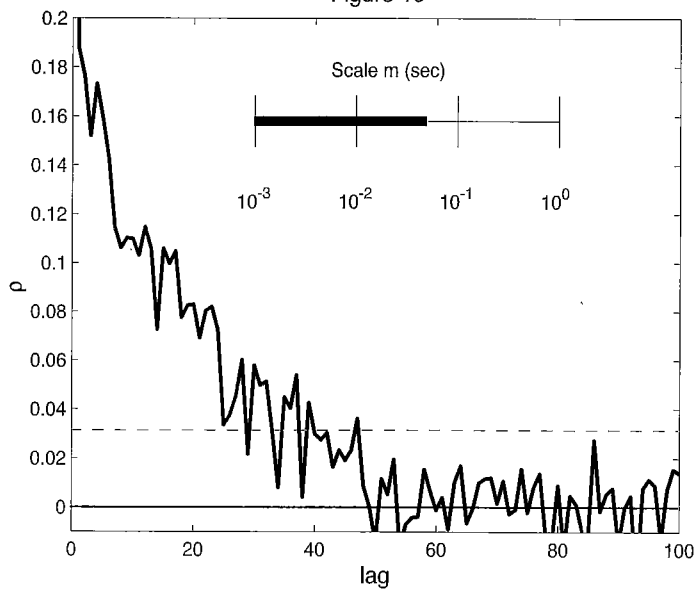

Figure 1c: Intermediate scale, $m = 0.05$, autocorrelation plot of 3 minute Internet traffic data set. Shows positive autocorrelation moving towards the left.

Figure 1c show another intermediate scale of $m = 0.05$ sec. At this scale the autocorrelation looks similar to what might be expected from say a standard ARMA process. Studying the transition, using the movie, from Figure 1b, it looks now more as expected, with the "dependence moving towards the right". Note also that at larger lags the long range dependence becomes obscured by the large sampling variability from too few terms in the autocovariance estimate. Finally observe that the red line has increased in height, because at this scale there are fewer bins, resulting in noisier covariance estimates.

Figure 1d shows the autocorrelation at scale $m = 1$ sec. At this scale there are only about 300 bins, so the autocorrelation function is severely subject to sampling variability. The absence of the dashed red line shows that the Gaussian White Noise quantile is completely above the range shown. Because of the large apparent variability, little reliance should be placed in these results.

Hence, the larger data set was gathered, as indicated in Section 2. A similar analysis to that of Figure 1 is shown in Figures 2a-d. Again, viewing the movie version of this, available in the file ZoomStatFig2.mpg in the same web directory, is highly recommended.

Figure 2a shows the smallest scale $m = 0.0003$ sec. Note that this autocorrelation function appears to be very similar to that of a Gaussian White Noise. In particular, roughly 5 percent of the values cross above the red dashed line. The reason this is noticeably different from Figure 1a, is that a longer series of bin counts was used for Figure1.

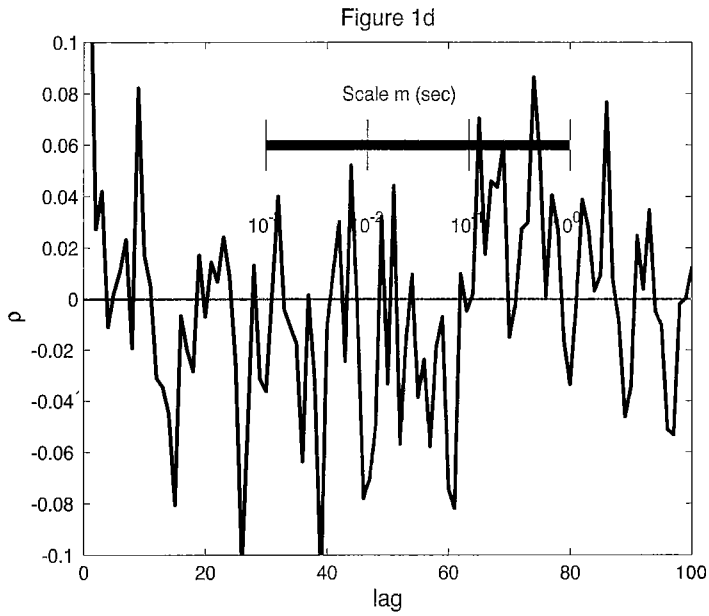

Figure 1d: Large scale, $m = 1$, autocorrelation plot of 3 minute Internet traffic data set. Shows large sampling variability.

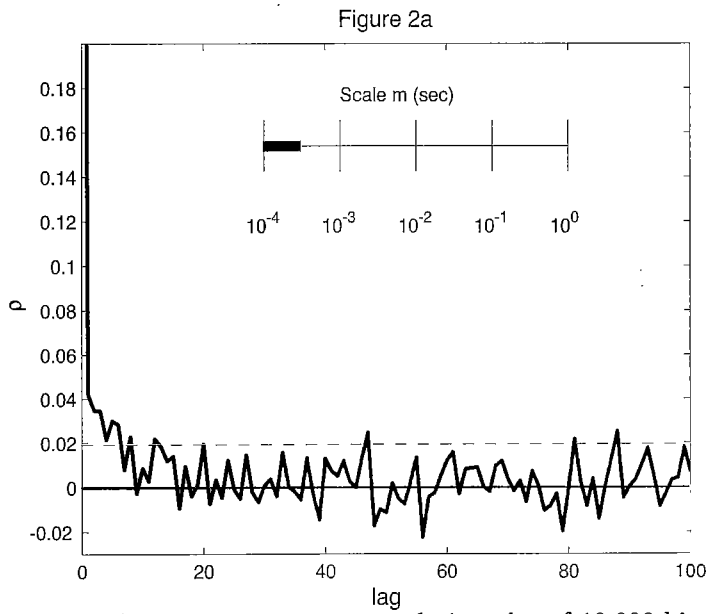

Figure 2a: Smallest scale, $m = 0.0003$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Shows autocorrelations very similar to a Gaussian White Noise process.

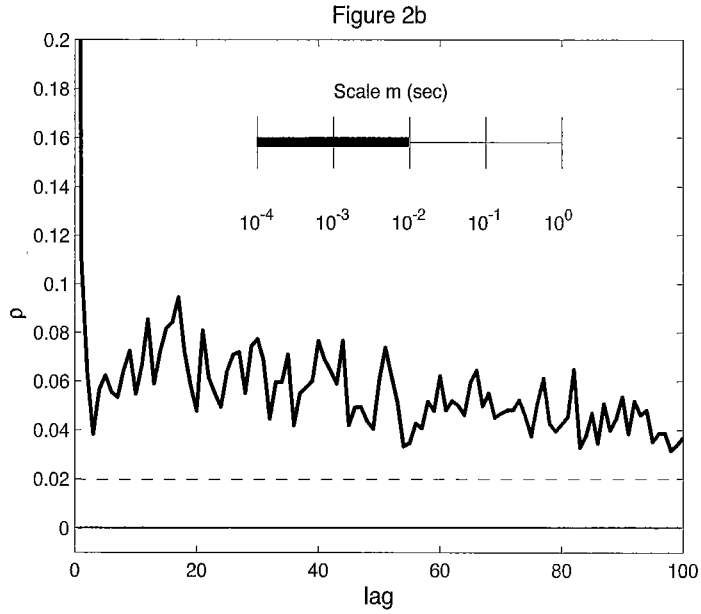

Figure 2b: Medium scale, $m = 0.01$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Shows strong positive autocorrelation.

Figure 2b shows the scale $m = 0.01$ sec. Here there is strong dependence at all lags shown. As for Figure 1b, the increasing correlation from Figure 2a does not "move in from the right", but instead "rises up vertically".

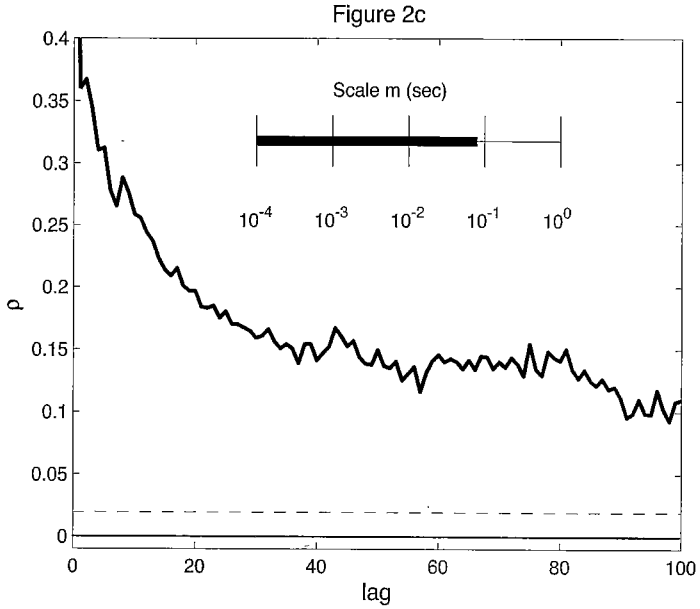

Figure 2c: Medium scale, $m = 0.08$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Shows strong positive autocorrelations.

Figure 2c shows the intermediate scale $m = 0.08$ sec. This is similar to Figure 1c for small lags but the positive correlation is much higher here for larger lags. This suggests that the larger lag structure in Figure 1c was driven by the data being too sparse.

Figure 2d

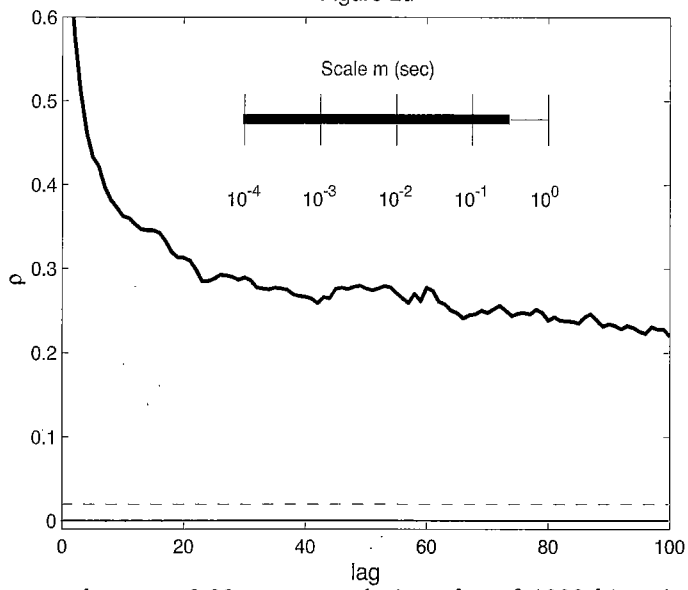

Figure 2d: Large scale, $m = 0.32$, autocorrelation plot of 1000 bin window, from one hour Internet traffic data set. Shows even stronger positive autocorrelation.

Figure 3a

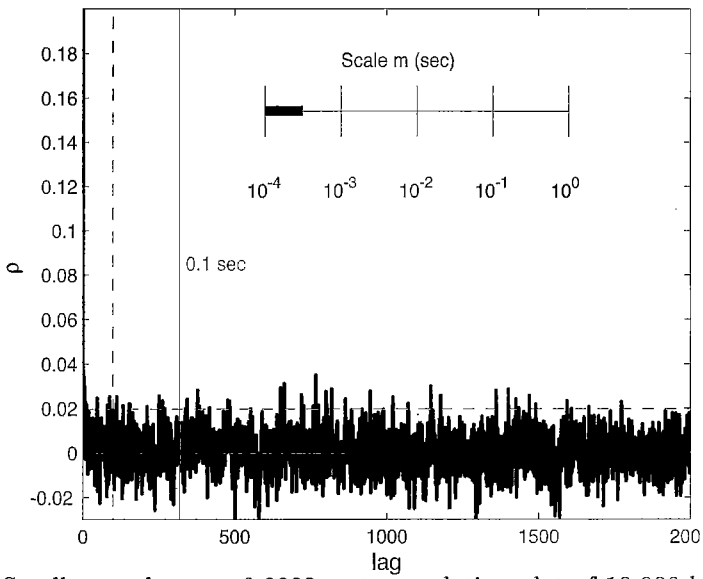

Figure 3a: Smallest scale, $m = 0.0003$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Shows autocorrelations very similar to a Gaussian White Noise process.

Figure 2d, at the scale $m = 0.32$ sec, suggests that the apparent white noise like structure in Figure 1d was probably due to working with a too small time series. Instead there is very strong positive correlation present for all lags shown at this scale.

A possible explanation for the surprising behavior in the small to medium scale transition, i.e. Figures 1a to 1b and Figures 2a to 2b, could be that too small a range of lags is considered. These pictures are reproduced, now with 2000 lags (vs. the 100 shown in Figures 2a-d) in Figures 3a-d. The vertical dashed line shows the previous upper limit of 100 lags. The "time changing intuition" of changing scales is also useful to view. For this some vertical red solid lines have been added to indicate several time scales (which is especially helpful in the movie version).

Figure 3a shows that the pattern of Figure 2a extends to a much wider range of lags. Here (and in Figures 3b - 3d) the vertical dashed blue line shows the corresponding boundary of Figure 2a. Again, this appears similar to the structure expected from white noise data, in that about 5% of the values cross above the dashed red line. This fits well with the structure found by Cao, Cleveland, Lin, and Sun (2001).

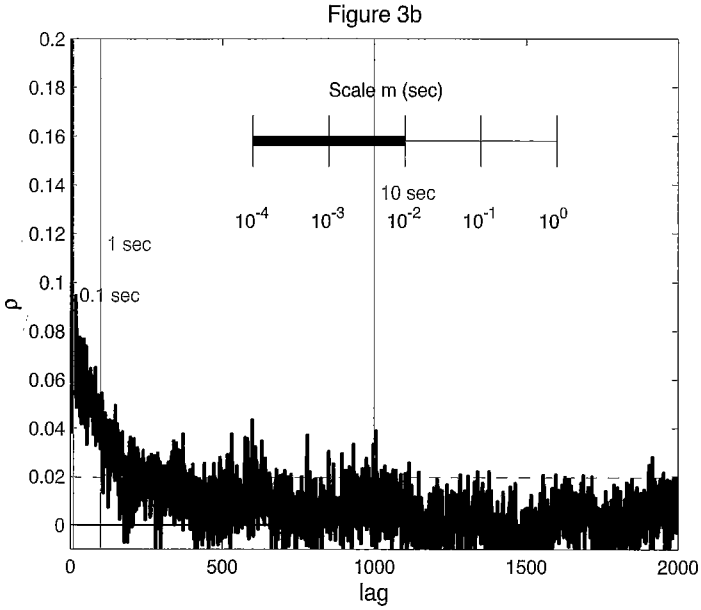

Figure 3b: Medium scale, $m = 0.01$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Shows strong positive autocorrelations fall off for larger lags.

Figure 3b shows that the strong positive correlation behavior, present in Figure 2b, falls off for larger lags. However, the movie version shows that the transition is still "vertical rise", not the "coming in from the right" that is expected from the usual internet traffic scaling observations which report strong correlations especially at round trip times.

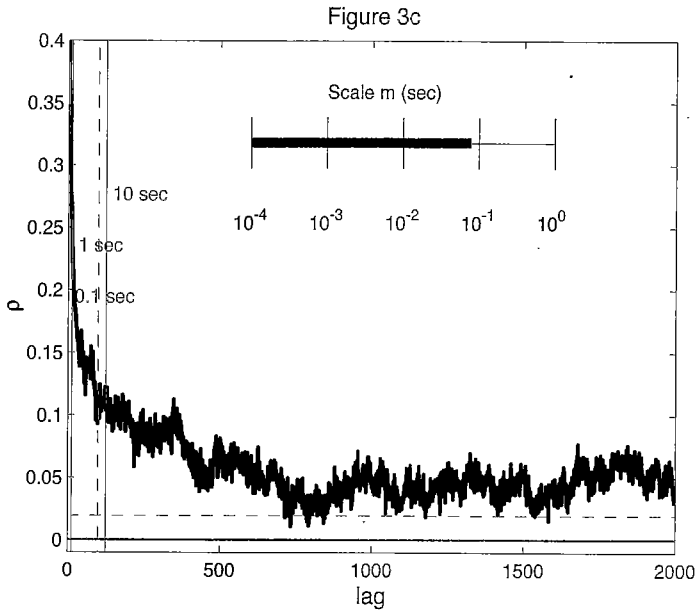

Figure 3c: Medium scale, $m = 0.08$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Continues to show strong positive autocorrelations falling off for larger lags.

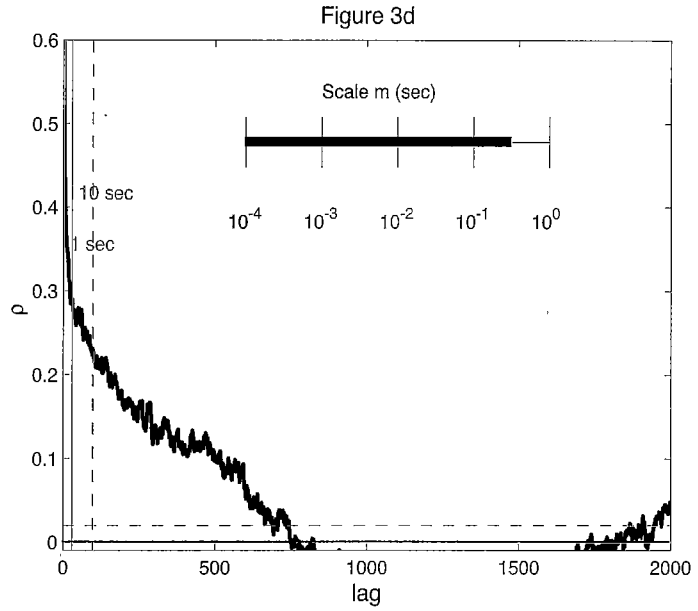

Figure 3d: Medium scale, $m = 0.32$, autocorrelation plot of 10,000 bin window, from one hour Internet traffic data set. Continues to show strong positive autocorrelations falling off for larger lags.

Figure 3c is for the scale $m = 0.08$ sec. Again for very large lags, the positive dependence falls off. The movie shows that for this transition, there is a definite impression of "dependence coming in from the right side" (e.g. focus carefully on the "10 sec" line).

Figure 3d continues the trend shown for Figure 3c. Studying the movie continues to show that in the transition, there is definite "right to left movement of the dependence".

In summary, the large scale behavior appears to correspond to the usual intuition (of Internet traffic researchers) that long range dependence happens at large scale, but the small scale behavior is quite different.

Some simple calculations can be used to understand this. Let $X_{2t}^{(m)}$ denote the $2t$-th bincount at scale m. When we shift to the coarser scale $2m$, this bincount (at the same time location) is replaced by

$$X_t^{(2m)} = X_{2t}^{(m)} + X_{2t+1}^{(m)}.$$

The corresponding autocovariance becomes

$$\begin{aligned} cov\left(X_{\bullet}^{(2m)}, X_{\bullet-l}^{(2m)}\right) &= cov\left(X_{2\cdot}^{(m)} + X_{2\cdot+1}^{(m)}, X_{2\cdot-2l}^{(m)} + X_{2\cdot-2l+1}^{(m)}\right) = \\ &= cov\left(X_{2\cdot}^{(m)}, X_{2\cdot-2l}^{(m)}\right) + cov\left(X_{2\cdot}^{(m)}, X_{2\cdot-2l+1}^{(m)}\right) + \\ &\quad + cov\left(X_{2\cdot+1}^{(m)}, X_{2\cdot-2l}^{(m)}\right) + cov\left(X_{2\cdot+1}^{(m)}, X_{2\cdot-2l+1}^{(m)}\right), \end{aligned}$$

where, letting n_m denote the number of bins at scale m (so that $n_{2m} = n_m/2$),

$$cov\left(X_{\bullet}^{(2m)}, X_{\bullet-l}^{(2m)}\right) = \frac{1}{n_{2m}} \sum_{t=l}^{n_{2m}} \left(X_t^{(2m)} - \overline{X}^{(2m)}\right)\left(X_{t-l}^{(2m)} - \overline{X}^{(2m)}\right),$$

$$cov\left(X_{2\bullet}^{(m)}, X_{2\bullet-k}^{(m)}\right) = \frac{1}{n_{2m}} \sum_{t=l}^{n_{2m}} \left(X_{2t}^{(m)} - \overline{X}^{(m)}\right)\left(X_{2t-k}^{(m)} - \overline{X}^{(m)}\right),$$

$$cov\left(X_{2\bullet+1}^{(m)}, X_{2\bullet+1-k}^{(m)}\right) = \frac{1}{n_{2m}} \sum_{t=l}^{n_{2m}} \left(X_{2t+1}^{(m)} - \overline{X}^{(m)}\right)\left(X_{2t+1-k}^{(m)} - \overline{X}^{(m)}\right),$$

for $k = 2l-1,\ 2l,\ 2l+1$, where $\overline{X}^{(m)}$ is the mean at scale m, with $\overline{X}^{(2m)} = 2\overline{X}^{(m)}$. Using the change of index $t' = 2t$, note that the second summation is over even $t' = 2, ..., n_m$. Similarly, using the change of index $t' = 2t+1$, the third summation is over odd $t' = 1, ..., n_m$. Thus putting both together gives the scale m covariance,

$$cov\left(X_{2\cdot}^{(m)}, X_{2\cdot-2l}^{(m)}\right) + cov\left(X_{2\cdot+1}^{(m)}, X_{2\cdot+1-2l}^{(m)}\right) = 2cov\left(X_{\cdot}^{(m)}, X_{\cdot-l}^{(m)}\right),$$

where the factor of 2 comes from $\frac{1}{n_{2m}} = \frac{2}{n_m}$.. The remaining two covariance terms are offset by ± 1. So when the autocorrelation function $\rho^{(m)}(l)$ is "smooth", i.e. doesn't change rapidly in l, (usually true in the above Figures), then the sum of the remaining terms is well approximated by

$$cov\left(X_{2\cdot}^{(m)}, X_{2\cdot-2l+1}^{(m)}\right) + cov\left(X_{2\cdot+1}^{(m)}, X_{2\cdot+1-2l-1}^{(m)}\right) \approx 2cov\left(X_{\cdot}^{(m)}, X_{\cdot-l}^{(m)}\right).$$

This "smoothness of the autocorrelation" is typical of most of the classical linear time series models such as ARMA (Auto Regressive Moving Average) processes, see e.g. Brockwell and Davis (1991). It follows that

$$cov\left(X_{\bullet}^{(2m)}, X_{\bullet-l}^{(2m)}\right) \approx 4cov\left(X_{\cdot}^{(m)}, X_{\cdot-l}^{(m)}\right).$$

Using a similar analysis, the scale $2m$ variance can be written in terms of scale m quantities as

$$\begin{aligned} var\left(X_{\bullet}^{(2m)}\right) &= cov\left(X_{2\cdot}^{(m)} + X_{2\cdot+1}^{(m)}, X_{2\cdot}^{(m)} + X_{2\cdot+1}^{(m)}\right) \\ &= var\left(X_{2\cdot}^{(m)}\right) + 2cov\left(X_{2\cdot}^{(m)}, X_{2\cdot+1}^{(m)}\right) + var\left(X_{2\cdot+1}^{(m)}\right) = \\ &= 2\left[var\left(X_{\cdot}^{(m)}\right) + cov\left(X_{\cdot}^{(m)}, X_{\cdot+1}^{(m)}\right)\right]. \end{aligned}$$

Combining the above results, we see that the autocorrelation scales approximately as:

$$\rho^{(2m)}(l) \approx \frac{4cov\left(X_{\cdot}^{(m)}, X_{\cdot-l}^{(m)}\right)}{2\left[var\left(X_{\cdot}^{(m)}\right) + cov\left(X_{\cdot}^{(m)}, X_{\cdot+1}^{(m)}\right)\right]} = \frac{2\rho^{(m)}(l)}{1+\rho^{(m)}(1)} = \frac{2}{1+\rho^{(m)}(1)}\rho^{(m)}(l). \tag{3.1}$$

A first consequence of (3.1) is that the autocorrelation scales approximately like a constant multiple over m. This explains the "constant horizontal increase" that was puzzling in the transitions from Figures 1a to 1b and from Figures 2a to 2b. Essentially the "small but positive and constant" autocorrelations are all scaled in nearly the same way. A second consequence of (3.1) is that the lag one autocorrelation, $\rho^{(m)}(1)$, is critical to the cross scale behavior of the autocorrelation. If this is quite small (e.g. in figures 1a and 2a), then the ratio $\frac{2}{1+\rho^{(m)}(1)} \approx 2$, so this scaling effect is quite large, which explains the rapid increase observed in the movie in those regions. On the other hand, when $\rho^{(m)}(1)$ is close to 1, e.g. as in Figures 2c and 2d, the ratio $\frac{2}{1+\rho^{(m)}(1)} \approx 1$, so this effect is negligible. In the latter situation, other considerations, such as "dependence coming in from the right" drive what is seen in the movie.

A challenging and interesting problem for future research is to find stochastic models with this type of scaling dependence structure. Classical ARMA models are inappropriate, since they have exponentially decreasing tails, while some "nearly constant, but small" correlation is needed for this type of behavior. The Poisson process structure suggested by Cao, Cleveland, Lin, and Sun (2001) for small scales also will not scale up appropriately, since in that case $\rho^{(m)}(l) \approx 0$, so even though the ratio $\frac{2}{1+\rho^{(m)}(1)} \approx 2$, 2 times 0 is still 0. While the autocorrelations in Figure 2a appear to be essentially zero, a simple hypothesis test shows that in fact the mean of the autocorrelations is significantly positive. This slight positivity is what drives the increasing trend seen in the transition to Figure 2b, as predicted by (3.1).

Another viewpoint for zooming autocovariances: instead of fixing the lag l, as the scale changes, fix time. In particular rescale the x-axis in the Figure 3 plots so that the

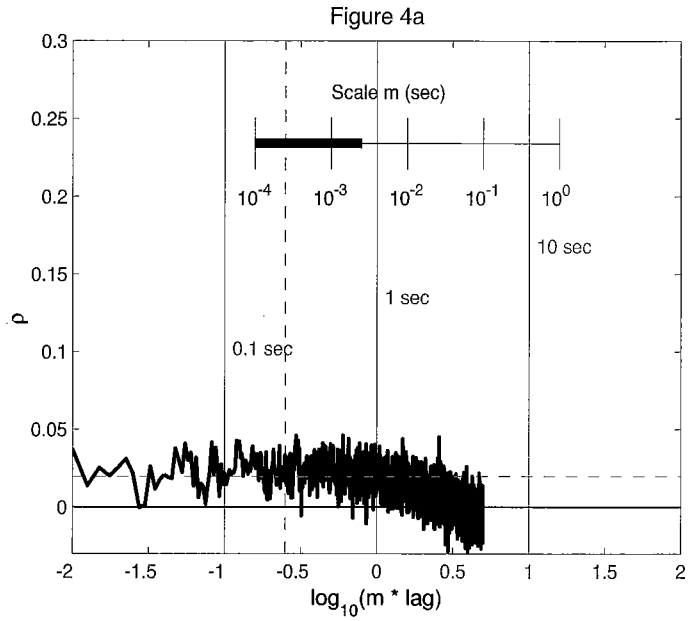

Figure 4a: Time fixed zooming autocorrelation of internet data. This again shows that for small scales the autocorrelation is not far from white noise.

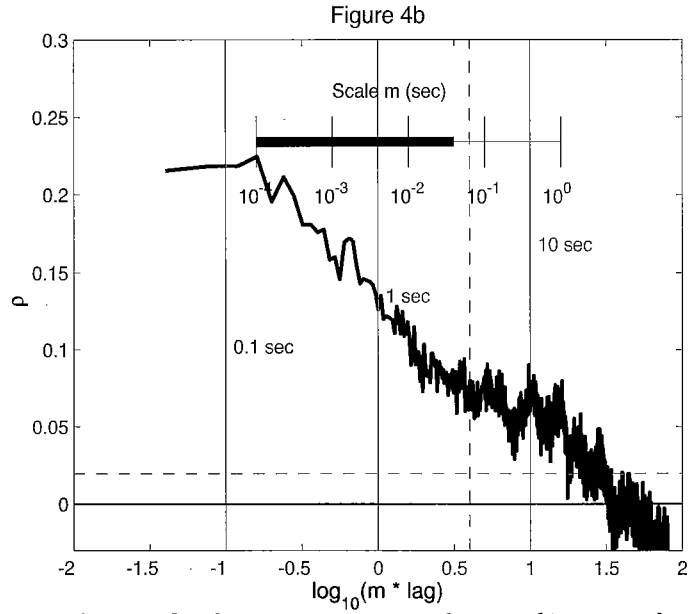

Figure 4b: Larger scale time fixed zooming autocorrelation of internet data. This shows dependence "rises", instead of "coming in from the right".

time points, represented by the solid vertical lines, remain constant. The results of this appear in Figure 4. Again the movie version in the file ZoomStatFig4.mpg is recommended. A smaller scale frame of that movie, for $m = 0.025$, is shown in Figure

4a. The same important times are highlighted with vertical red lines as in Figures 3a-d. Note that now a logarithmic time scale is used.

This shows dependence similar to that demonstrated above. But the importance of this rescaling comes from looking across scales, as in the movie. A larger scale, $m = 0.04$ frame is shown in Figure 4b. Note that in Figure 4b, the time axis, and the vertical red lines stay in the same position.

Figure 4b shows how the above described "lifting of the dependence" appears on this scale. This, and especially the movie, show that on this scale the dependence "moves towards the right". This again is contrary to the usual internet traffic intuition of "dependence existing at certain scales". Under that model, the heights should stay roughly constant in this view of the autocorrelations.

An interesting open problem is to find effective models for this unusual long range dependence. This analysis makes it clear that at small scales a very small, but positive (and nearly constant) autocorrelation is needed. Then aggregation will give appropriate auto-correlation structure.

4. Zooming SiZer

SiZer (Significance of Zero crossings of the derivative), is an exploratory data analysis tool proposed by Chaudhuri and Marron (1999). A straightforward application of SiZer, for the same data set as in Figure 1, is shown in Figure 5a. As for the above figures, Figure 5a is one frame of a movie. This movie is in the file ZoomStatFig5.mpg, in the same web directory, and its viewing while reading this is strongly recommended.

The "raw data" that is input to SiZer in Figure 5a, is bin counts, from 400 equally spaced bins, over the full time range (about 180 secs). The top figure is a family of local linear smooths, see e.g. Fan and Gijbels(1996) or Wand and Jones (1994) for access to the literature on this method. The family of smooths uses the range of bandwidths shown as the y-axis in the bottom panel, $h \in [-0.4, 1.3]$. In the top panel, there does not seem to be any important structure, or trends in the data. However the bottom panel shows that in fact there is a great deal of structure that can not be explained as "random noise". The bottom panel is the SiZer map, which assesses the statistical significance of the slope of the smooth, for each time (shown on the horizontal axis) and for each bandwidth h, (shown on the vertical axis). At locations where the slope is significantly increasing (i.e. the hypothesis of 0 slope can be rejected in this direction, i.e. a confidence for the slope is completely above 0), the color blue is used. Similarly the color red is used in areas of significant decrease. The intermediate color of purple is used in locations where there is no significant slope (i.e. a confidence interval for the slope contains 0). The dashed white lines give an intuitive idea of the amount of smoothing being done, representing the "effective width of the Gaussian window function" by the center plus and minus two standard deviations.

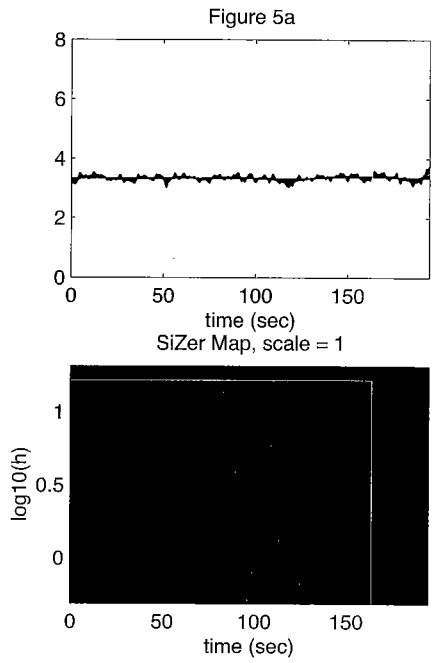

Figure 5a: Largest scale (convential) SiZer analysis of the 3 minute Internet traffic data set. Shows rich structure, caused by the "burstiness" of internet traffic data.

If these data came from a homogenous Poisson process, so the wiggles observed in the top panel were just spurious sampling noise artifacts, then the SiZer map would be entirely purple. The large number of red and blue regions that are visible here indicate the existence of strongly changing trends in this time series, that appear at a wide range of scales. This is a new view of the "burstiness" that has been found by others. A "burst of data" passing through the link appears as blue near the beginning of the burst (when the traffic level is sharply increasing) followed by red at the end of the burst (when the traffic level tapers off).

Note that the significant red and blue regions appear even at the smallest scales, i.e. the bottom of the SiZer map. This suggests that there may be interesting structure at finer scales, which is not surprising since there are 1 million data points, and only 400 bins. Zooming SiZer provides a cross-scale visualization that shows clearly and continuously how the SiZer map changes across a wide range of scales.

Zooming SiZer is a sequence of pictures of the type shown in Figure 5a, where in each succeeding picture, only about the first 84% of the data in the previous picture is used. The scale factor is actually $2^{-1/4} \approx 0.84$ so that 4 steps gives a factor of 2. The yellow line in the top panel of Figure 5a shows this boundary (for the next picture). Since a smaller bandwidth range is used in the next picture, the bandwidth range is reduced by the same factor. This results in a new SiZer map, where the upper and right hand boundaries are marked in yellow on the lower panel of Figure 5a, together with some

new area at the bottom (to give a corresponding bandwidth range). The next picture in the series is shown in Figure 5b.

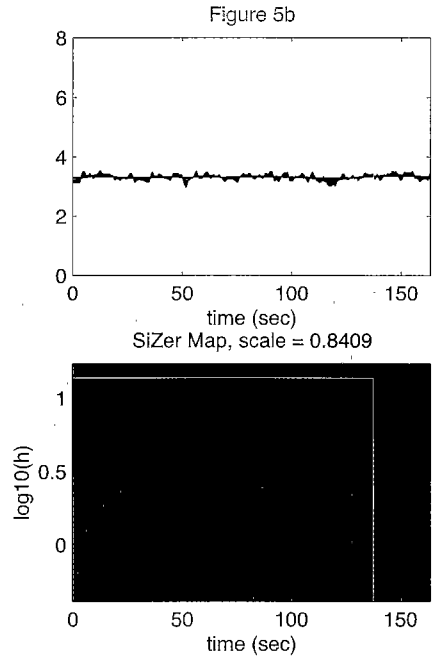

Figure 5b: Next smaller scale in the zooming SiZer analysis of the 3 minute Internet traffic data set. This represents the area inside the yellow box in Figure 5a, plus some new area at the bottom.

Note that the overall pattern in Figure 5b is very similar to that inside the yellow box in Figure 5a, except near the edges where boundary effects dominate. At this point it is quite helpful to download that movie, and view it in that way. The reason is that the red and blue regions move nicely as the zooming SiZer progresses down through the scales. A much finer scale is shown in Figure 5c.

Here the scale is an order of magnitude smaller. There is much more purple on this SiZer map, indicating that most of the structure apparent in the family of smooths in the top panel could be explained by random variability. However, there are still a few regions of significant change, indicating that also at this time scale, there occasional bursts in the data.

The finest scale considered here is shown in Figure 5d.

Here there is a gray fringe appearing at the bottom. SiZer uses this color to indicate regions where there is not enough data to do reliable statistical inference. Note that even at this much smaller time scale (in particular three orders of magnitude smaller than that of Figure 5a), there are still occasional significantly bursty locations.

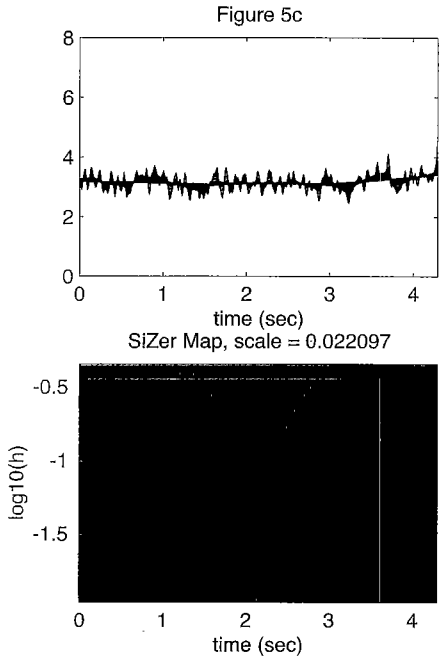

Figure 5c: smaller scale part of the zooming SiZer analysis of the 3 minute Internet traffic data set. Shows a smaller amount of structure that can be distinguished from random.

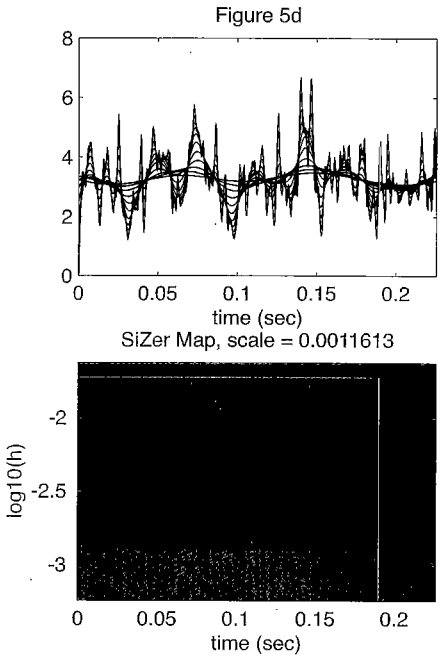

Figure 5d: Smallest scale in the zooming SiZer analysis of the 3 minute Internet traffic data set. The gray represents not enough data for reliable analysis. This shows nonrandom structure, even at the smallest scale.

5. Acknowledgments

This research was supported by NSF Grant DMS-9971649. The authors are grateful to F. D. Smith for providing the internet data.

REFERENCES

brockwell91 Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, Springer Verlag, New York.

chaudhuri99 Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807-823.

fan94 Fan, J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, 3, 35-56.

fan96 Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

feldman98 Feldmann, A. Gilbert, A. C. and Willinger, W. (1998) Data networks as cascades: investigating the multifractal nature of Internet WAN traffic, *Computer Communication Review, Proceedings of the ACM/SIGCOMM '98*, 28, 42-55.

paxson95 Paxson, V. and Floyd, S. (1995) Wide Area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, 3, 226-244.

percival00 Percival, D. B. and Walden, A. T. (2000) *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge.

riedi99a Riedi, R., Crouse, M. S., Ribeiro, V. and Baraniuk, R. G. (1999) A Multifractal Wavelet Model with Application to TCP Network Traffic, *IEEE Transactions on Information Theory*, 45, 992-1018.

riedi99b Riedi, R. and Willinger, W. (1999) Toward and improved understanding of network traffic dynamics, in *Self-similar Network Traffic and Performance Evaluation*, Wiley, New York.

wand96 Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, London.