A SIZER ANALYSIS OF IP FLOW START TIMES

J. S. MARRON, FELIX HERNANDEZ-CAMPOS, AND F. D. SMITH

ABSTRACT. The SiZer technique is used to study the homogeneity of a point process of Internet traffic flow start times. It is seen that a homogenous Poisson process is an inappropriate model, because it does not yield observed statistically significant burstiness. Some Weibull waiting processes gives better, but still inadequate performance. A clustered Poisson process gives the best fit.

1. INTRODUCTION

Simulation of Internet traffic is a challenging and interesting problem. It is important for both Internet researchers (who try to improve the performance of the Internet itself), and also for testing of many types of Internet based business applications. There is a strong need for involvement of creative statisticians, both in the development of the simulation methods, and in the assessment of their performance. This paper discusses an example demonstrating the usefulness of the relatively new statistical method, SiZer, and showing that simulation model development and statistical model assessment can and should interact.

Most of Internet traffic is composed of IP (Internet Protocol) flows. These are the transfers of data from one computer to another, as described in RFC 791, see Postel (1981). An IP flow is defined herein as a set of packets carrying IP datagrams that share the same sending and receiving addresses (more specifically, that have identical IP addressing 5-tuples consisting of protocol number, source IP address and port number, and destination IP address and port number). Transport protocols, such as TCP (Transmission Control Protocol) and UDP (User Datagram Protocol), provide higher-level communication services built on top of IP that support the exchange of information between applications. Web browsing, email, telnet and many others use TCP over IP, while audio/video streaming, name resolution, and other applications use UDP over IP. IP is therefore the fundamental building block of the Internet, and its complex behavior is the result of aggregating the communication patterns of very

diverse applications. Consequently, modeling IP is a remarkably challenging problem that has attracted a lot of attention over the last decade.

One way to simulate overall traffic is to simulate the flows, and then to aggregate them. There are three essential components of this, the point process of flow starts, the duration distribution of the flow, and the structure of the transmission within the flow. In this paper only the point process of the flow start times is considered. See Paxson (1994), Danzig, Jamin, Caceres, Mitzel, and Estrin (1992), Cleveland, Lin and Sun (2000), Feldmann (2000) and Cao, Cleveland, Lin and Sun (2001) for important work on this. See Garrett and Willinger (1994), Leland, Taqqu, Willinger, and Wilson (1994) and Paxson and Floyd (1995) and Crovella and Bestavros (1996) for access to the large literature on flow duration distributions, and some very interesting implications of that work. See Kulkarni, Marron and Smith (2000) for an approach to simulation of within flow traffic structure.

Figure 1 is a real data visualization of the point process of the start times of $n = 115548$ IP flows. Each dot represents one flow start, with the x coordinate representing the start time, and a "jittered" random y coordinate to separate the dots for easy viewing. Because showing all of the data would result in massive overplotting, only a random sample of 2000 is shown here. All data sets in this paper are one dimensional (events occurring in time), with some vertical jittering (see e.g. pages 121-122 of Cleveland 1993) occasionally added just for visualization. The data were collected at the main link between the University of North Carolina at Chapel Hill and the rest of the Internet, on a Sunday morning in 2000. The methodology used in this traffic capturing effort is described in Smith, Hernandez-Campos, Jeffay and Ott (2001). Here all packets in a time interval of approximately 40 minutes are considered, and a dot is shown the first time a new pair of sending and receiving address is observed. To avoid the boundary effect of dots piling up on the left end (indicating flows extending beyond the left in an unobservable way), the left 20% of the picture is not shown. Using this visualization, it is hard to notice structure in the data that is different from a random uniform distribution of the points.

A simple model for random uniformly distributed events, such flow start times as shown in Figure 1, is a Poisson point process. See any standard probability text covering point processes, e.g. Chapter 3 of Resnick (1987), for a detailed description

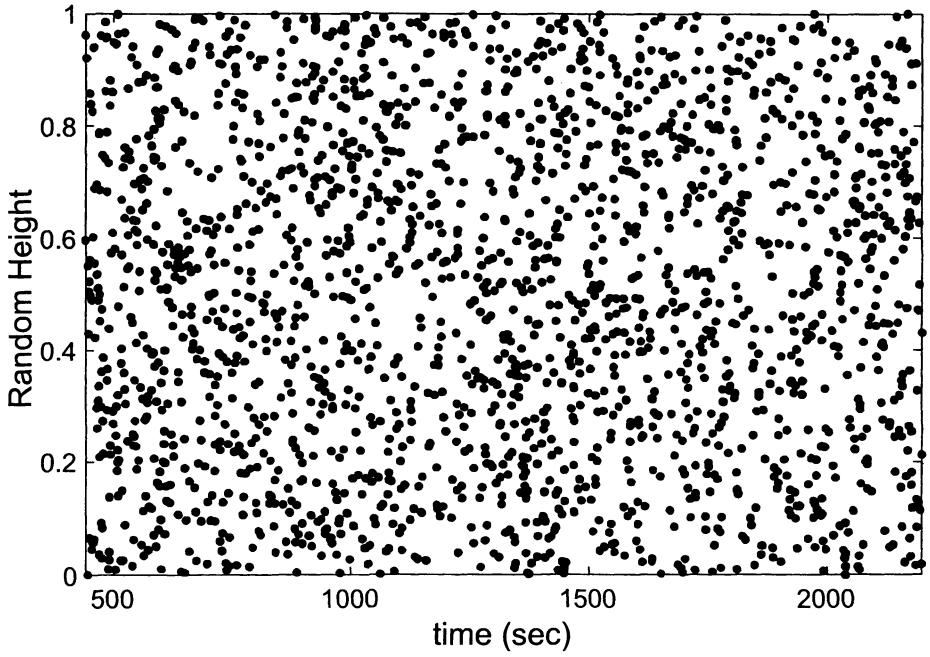

FIGURE 1. *Visualization of IP flow start times, from main link between UNC and the Internet. Random sample of 2000 from 111822.*

and development. Such a model is characterized by independent exponential inter-arrival times (time spacing between consecutive dots). If the data are binned to any equally spaced grid, the bin counts are independent, and have a Poisson distribution with the same mean parameter over different bins. This mean parameter is called the "intensity", and the full process is called a "Homogeneous Poisson Process". A very relevant variation is the "Nonhomogeneous Poisson Process", where the intensity is a function of location, i.e. the points are more dense in some locations and more sparse in others.

The goal of this paper is the careful investigation of the time varying intensity function of the point process shown in Figure 1, and the intensity of some simulation models. This function can be estimated using essentially kernel density estimation methods. A critical question is whether or not "bumps" observable in the kernel smooth represent important departures from the standard homogeneity structure, or instead could be explained by the natural sampling variability inherent to the Poisson

Process. The SiZer method is a powerful and convenient visual device for answering this question.

A quick introduction to SiZer, together with an analysis of the data in Figure 1 is given in Section 2. It is seen that the data in Figure 1 is clearly not a homogeneous Poisson model. In particular, the data exhibit more "clumping" than a standard homogeneous Poisson Process (i.e. a tendency towards "more than random clustering"). The SiZer analysis shows that this difference is statistically significant. This motivates a search for a start time model with characteristics closer to the data.

Cleveland, Lin and Sun (2000) and Cao, Cleveland, Lin and Sun (2001) have suggested that inter-arrival times may be better fit by non-exponential Weibull distributions. They construct a start time model by combining Weibull interarrivals with an appropriate long range dependence structure.

In this paper, independent Weibull inter-arrival times (with shape parameter less than 1) are aggregated to give a non-Poisson start time process, that might be expected to yield clumping of the type found in Section 2. Such processes are considered in Section 3, where SiZer is used to understand the characteristics of two variations of this. Unfortunately, sensible estimates of the Weibull shape parameter do not give the right level of bumps. While bumpiness similar to that in the real data can be generated using a deliberately tuned Weibull shape parameter, it is seen that the inter-arrival times are then completely inappropriate.

This motivates an alternate model, the "Clustered Poisson Process", considered in Section 4. It is seen that an appropriately tuned version of this model has much better SiZer properties, that are much closer to those of the real data. Hence the Clustered Poisson Process is recommended for further investigation as a leading candidate for the simulation of IP flow start times. Moreover, these results highlight the importance of the inclusion of flow dependencies in traffic models. The behavior of some applications that drive Internet traffic provides an intuitive explanation of these dependencies. In particular, the start times of worldwide web flows naturally cluster around web pages.

2. Initial SiZer Analysis

SiZer (shortening of SIgnificance of ZERo crossings of the derivative) is a method for understanding statistically significant features in smoothing methods. These methods include scatterplot smoothing, i.e. nonparametric regression, and smoothed histograms, i.e. kernel density estimation. Poisson intensity estimation is very closely

related to kernel density estimation. In particular, Homogeneous Poisson Process data, conditioned on the sample size (the total number of points in the picture), have a uniform probability density. So finding nonhomogeneity of point process data is equivalent to finding non-constancy in the slope of the corresponding density.

A SiZer analysis of the data shown in Figure 1 appears in Figure 2. The top panel shows a number of blue curves, which are kernel density estimates, reflecting the local intensity (i.e. higher where there are more, lower where there are less) of the points displayed in Figure 1 (a subset of which are displayed as jittered green dots). There are several blue curves corresponding to different levels of "smoothing". The level of smoothing is roughly the binwidth of the underlying histograms, but more precisely is the "bandwidth", i.e. the standard deviation of the smoothing window functions.

Some of the blue curves appear quite smooth, suggesting clear homogeneity of the point process in Figure 1. These correspond to a large bandwidth. Others, corresponding to smaller bandwidths, suggest some regions of nonhomogeneity, i.e. "bumps" in the distribution of points (suggesting "burstiness") in Figure 1.

But do these bumps represent important underlying structure, i.e. genuine nonhomogeneity of the process, or are they simply artifacts of the natural Poisson variability? This important question is addressed using the SiZer method developed by Chaudhuri and Marron (1999). A detailed introduction, with examples, can be found at the web site:

`http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer_Intro.html`.

An important component of SiZer is the "scale space" view of smoothing, where a wide family of smooths, such as those in the top panel of Figure 2, is considered.

The density estimation version of the SiZer map, shown in the bottom panel of Figure 2, is based on confidence intervals (at significance level $\alpha = 0.05$) for the slopes of the kernel density estimates. Experimentation shows that changing this significance level over reasonable values ($\alpha = 0.001$ to $\alpha = 0.2$) results in generally small changes in the SiZer map The rows of the map indicate different scales (i.e. bandwidths), and each row corresponds to a blue curve shown in the top panel. When the confidence interval is completely above 0, the slope is significantly increasing, and the color blue is used. When it is completely below 0, it goes down, where red is used. When the confidence interval contains 0, the slope is unclear, and the color purple is used. Visual correspondence between the scales (bandwidths) and the inference being done

in the SiZer map is given by the dashed white curves. These show the "effective window width" of the smooths, in terms of plus or minus two standard deviations of the Gaussian window function at each level of resolution.

The confidence intervals used in SiZer are based on standard central limit theory calculations. This requires a reasonable number of data points in each window. This is not an issue in this paper, because of the large size of the data sets. For smaller data sets, the fourth color of gray is used in locations where there is not enough data for reliable inference.

An important issue is to make the confidence intervals simultaneous, i.e. to take the multiple comparison problem into account. This is done using an "effective number of independent blocks" adjustment.

The SiZer map in Figure 2 shows a large amount of nonhomogeneous structure in the data displayed in Figure 1. The red at the top shows that at larger scales there is statistically significant downward trend in start times. This is caused by the fact that simply cutting off the first 20% of the flows is only a crude boundary adjustment, and even over this range, flows are still a little more likely to start at the beginning than at the end. The effect is not strong, but with $n = 115548$ data points, SiZer will find even small departures from homogeneity.

The lower part of the SiZer map shows many red and blue regions. These correspond to "finer scale views of the data", i.e. the use of smaller bandwidths (essentially histogram binwidths). The red and blue colors show that the slopes of the blue curves, which make up the bumps in the top panel, are steeper than would be generated by a homogeneous Poisson process. In other words this point process is more bursty than data simulated from a naive Poisson model.

This idea is checked in Figure 3, where the SiZer analysis is repeated for data simulated from a homogeneous Poisson Process. These simulated data use the same time interval, and are conditioned on the same number $n = 115548$ of points. Note first that the family of blue curves in the top panel of Figure 3 include some very smooth (large scale) members, and some wiggly (fine scale) members. The magnitudes of the small scale wiggles appear to be perhaps smaller in Figure 3 than in Figure 2. The corresponding SiZer maps show that this difference is very marked in the sense of statistical significance. The Figure 3 SiZer is almost completely purple, indicating that

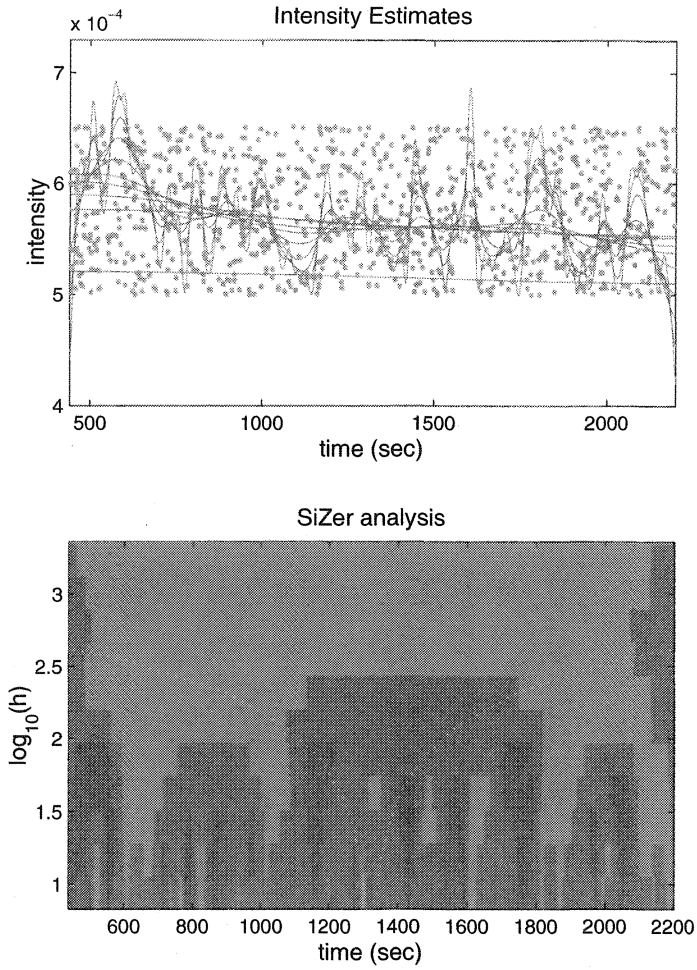

FIGURE 2. *SiZer analysis of the full data set from Figure 1. The family of intensity estimates appears in top panel, and SiZer map in the bottom. The SiZer map reveals statistically significant burstiness.*

the wiggles in the top panel are no larger than expected due to the natural variability in a homogeneous Poisson process.

There are some blue and red regions in the SiZer map, that are caused by boundary effects. In all the SiZer maps shown in this paper, boundary effects are mitigated using a simple "reflection" adjustment, where the data are reflected beyond the endpoints and included in the estimation process. See e.g. Section 2.10 of Silverman (1986) or

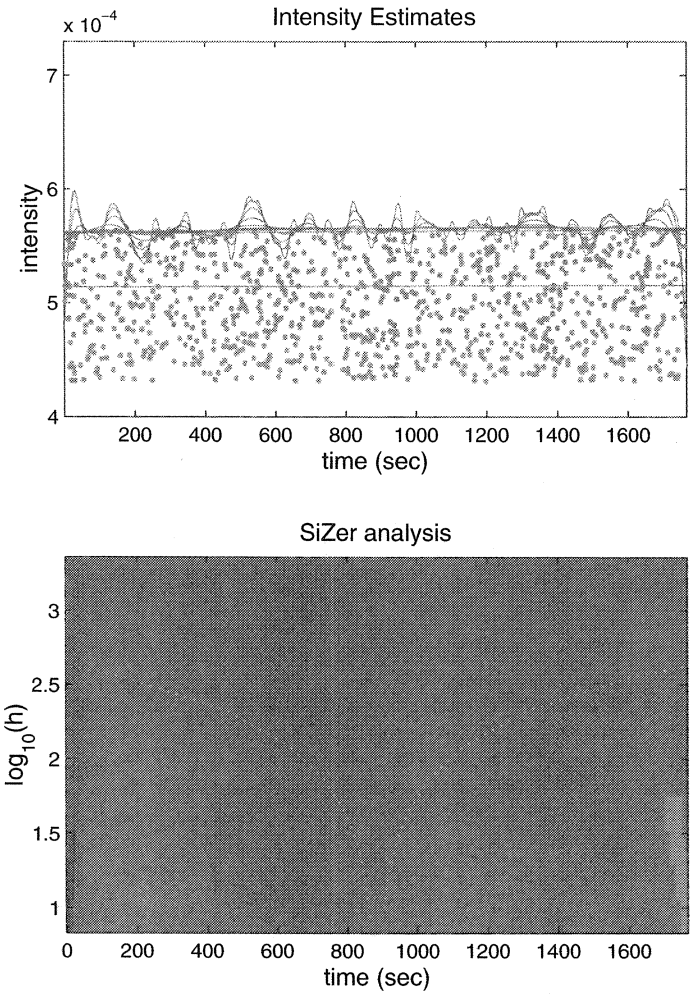

FIGURE 3. *SiZer analysis of simulated homogeneous Poisson process data, showing that SiZer does not label bursts here as statistically significant.*

Section 6.2.3.5 of Scott (1992) for description and further discussion. While this approach significantly decreases boundary effects, these show that they do not disappear completely.

3. A Weibull waiting time model

The lesson from Figure 2, that the point process of inter-arrival times is not exactly a Poisson process, has already been observed in other contexts. For example, Cleveland, Lin and Sun (2000) and Feldmann (2000) show that a Weibull distribution with a shape parameter less than one (thus not exponential, in the direction of "heavier tails") frequently provides a better fit to the distribution of packet inter-arrival times.

Heavier tailed inter-arrival time distributions are expected to produce "more than the usual burstiness" in the overall process, as observed in Figure 2, because there will be more longer and also more shorter interarrivals than expected for the exponential waiting time. This motivates a careful look at the inter-arrival time distribution.

The distribution of the inter-arrival times for the data of Figure 1 is analyzed in Figure 4A, using a Q-Q (Quantile vs. Quantile) plot. This is a graphical method for assessing the goodness of fit of the exponential distribution to the data. The main Q-Q plot is the red curve, which shows the data quantiles (i.e. the sorted data values) on the vertical axis, with the corresponding theoretical quantiles, from the exponential distribution, on the horizontal axis. When the theoretical distribution is a good fit to the data, the red curve should be "close" to the forty five degree line through the origin, shown here as a diagonal green line. A serious practical hurdle for this type of analysis, has been understanding what is meant by "close", because the red curve never completely follows the diagonal green line due to sampling variability. The envelope of blue curves provides a simple visual accounting for this natural variability. Each of these curves is a similar Q-Q plot, where the "data" are simulated from the theoretical distribution (with the same sample size of $n = 115548$), to reflect the variability inherent to the sampling process. Good visual impression comes from overlaying 100 such blue curves. When the theoretical distribution is a good fit to the data, the red curve should lie mostly within the blue envelope. The observed substantial departures of the red curve from the blue envelope indicate strong lack of fit between the data and the theoretical distribution.

In Figure 4A, the theoretical distribution (with quantiles shown on the horizontal axis) is the Exponential distribution (i.e. the Weibull distribution with shape parameter $\alpha = 1$), with scale parameter $\sigma = 0.0157$ (taken to be the sample mean, which is the maximum likelihood estimate).

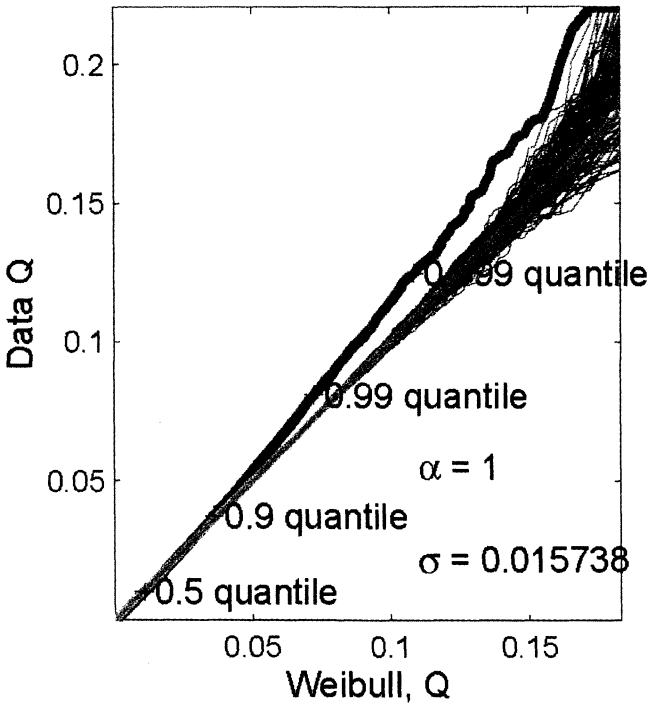

FIGURE 4. (A) *Q-Q analysis (red curve). of the data from Figure 1 compared to an Exponential distribution (with parameters shown). Blue envelope shows sample variability, making it clear that the Exponential distribution is a poor fit.*

Figure 4A shows that the Exponential distribution does not provide an acceptable fit to the inter-arrival times for the point process data shown in Figure 1, because the red curve is far outside the blue envelope in several places. Furthermore, it is seen that the real data have a "heavier tail" than the exponential because the red curve is well above the blue envelope for the larger quantiles. This is completely consistent with the lesson from Figure 2, that the original data are not well fit by a homogeneous Poisson process.

The Q-Q plot in Figure 4A also suggests that perhaps a fairly simple modification of the theoretical distribution could yield an appropriate fit. In particular, some modification of the Weibull shape parameter might work. This path is pursued in

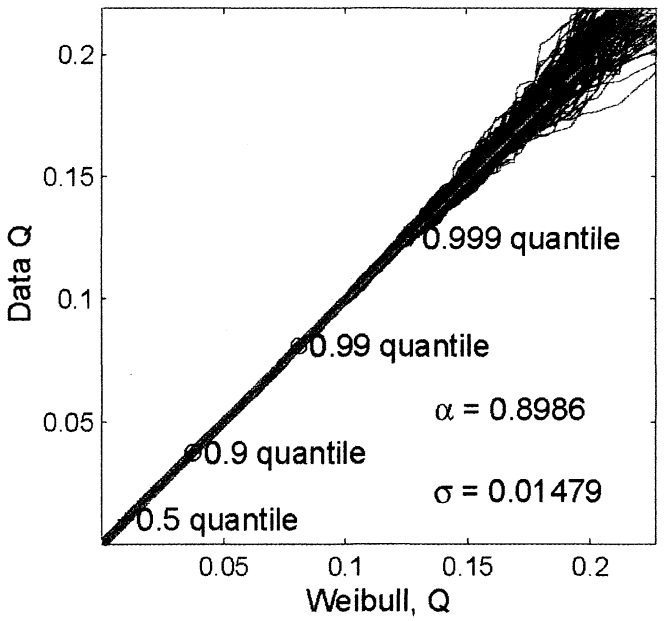

FIGURE 4. (B) *Q-Q analysis (red curve). of the data from Figure 1 compared to a Weibull distribution (with parameters shown, estimated by quantile matching). Blue envelope shows this Weibull distribution gives an acceptable fit.*

Figure 4B, where the theoretical underlying exponential distribution has been replaced by a more general Weibull distribution. This time the Weibull parameters α and σ have been estimated by "quantile matching", in particular they have been chosen to make the theoretical 0.99 and 0.999 quantiles the same as the data (i.e. the red curve crosses the green line at these two points indicated by circles in Figure 4B).

This Q-Q plot (the red curve) lies almost entirely within the blue envelope, suggesting that this Weibull distribution gives an acceptable fit to the data. The estimated shape parameter is $\alpha = 0.90$, which means a tail that is heavier than for the Exponential distribution, which appears to be consistent with the statistically significant burstiness observed using the SiZer method in Figure 2.

A natural way to use this information in an improved start time simulation model, is to use independent Weibull variables (with parameters $\sigma = 0.0148$ and $\alpha = 0.90$)

to generate the spacings (inter-arrival times) between the events in a point process. It is seen in Figure 4B that the Weibull distribution is correct, but the assumption of independence is not so clear. This issue is addressed in Figure 5, where a SiZer analysis is used to assess the burstiness of this independent Weibull(0.9) point process. As above $n = 115548$ data points were simulated. The total time span shown is the sum of the simulated realizations.

The magnitude of wiggles in the blue family of intensity estimates in the top panel looks more like the simulated data in Figure 3 than the real data in Figure 2. This impression is confirmed by the SiZer map in the bottom panel, which shows that none of the bursts are statistically significant. The SiZer map also shows a small amount of boundary effect, as in Figure 3. This shows that the point process in Figure 2 is clearly not an independent Weibull(0.9) inter-arrival process. While the Weibull(0.9) distribution is correct, the independence is not. Generating dependent processes is much more complicated (because a dependence structure needs to be specified). Some ideas for addressing this problem are given in Section 4.

A surprising feature in the SiZer map is the red band near the top. This indicates a statistically significant downward trend, which suggests that the heavy tail of the Weibull inter-arrival distribution creates some perhaps unexpected type of spurious dependence.

An interesting side issue is whether *any* independent Weibull inter-arrival time process could generate the type of burstiness observed in Figure 2. In particular, heavier tailed Weibull distributions should induce both longer and shorter waiting times, which should result in significant burstiness. Some experimentation with the Weibull shape parameter showed that $\alpha = 0.45$ was quite interesting in this respect. The resulting SiZer analysis is shown in Figure 6 (using the simulation scheme as for Figure 5, except that now $\alpha = 0.45$). Other values of α will be of interest to some, but are omitted here to save space. However, these can by viewed in the files UNC2000FlowSimWeibToy1IntArrs20.ps, UNC2000FlowSimWeibToy2IntArrs20.ps, UNC2000FlowSimWeibToy3IntArrs20.ps and UNC2000FlowSimWeibToy4IntArrs20.ps in the web directory `http://www.unc.edu/depts/statistics/postscript/papers/marron/NetworkData/StartTimeSiZer/`

The magnitude of the wiggles in the top panel of Figure 7 is similar to those in Figure 2, although the "frequency" may not be the same. The corresponding structure in the

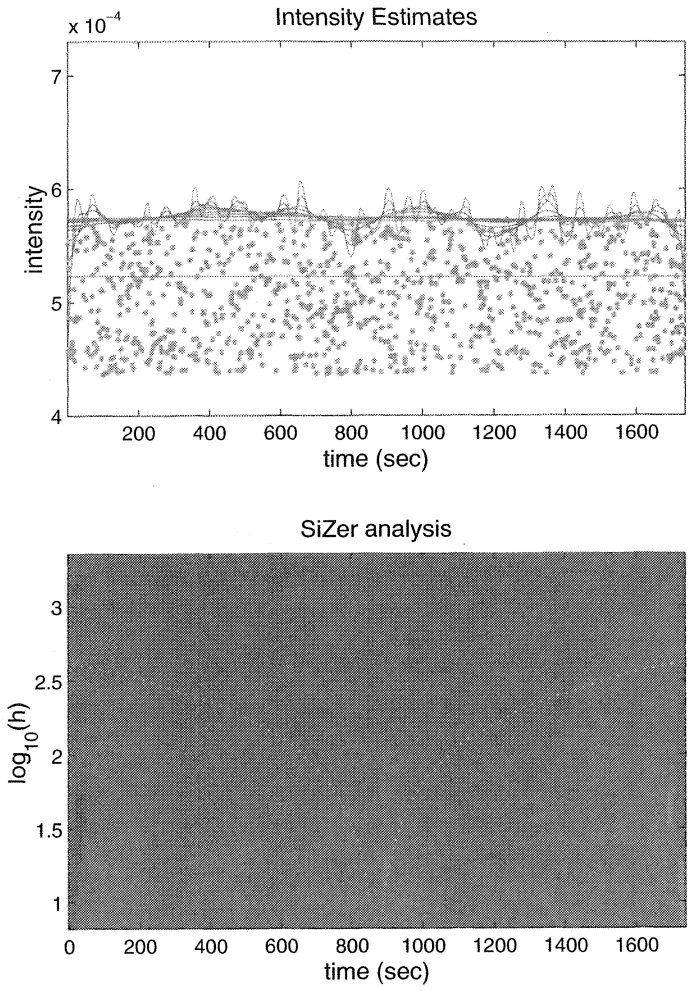

FIGURE 5. *SiZer analysis of start times simulated by Weibull(0.90) renewal model. Shows structure closer to homogeneous Poisson process, than to real data.*

SiZer maps are similar near the medium scales, i.e. window widths near $\log_{10} h \approx 1.5$. However, they are different at the finer scales, i.e. the smaller window widths. This shows that while the independent Weibull(0.45) inter-arrival process gets the medium scale burstiness approximately correct, it only does so at the cost of introducing some fine scale burstiness that is not present in the real data.

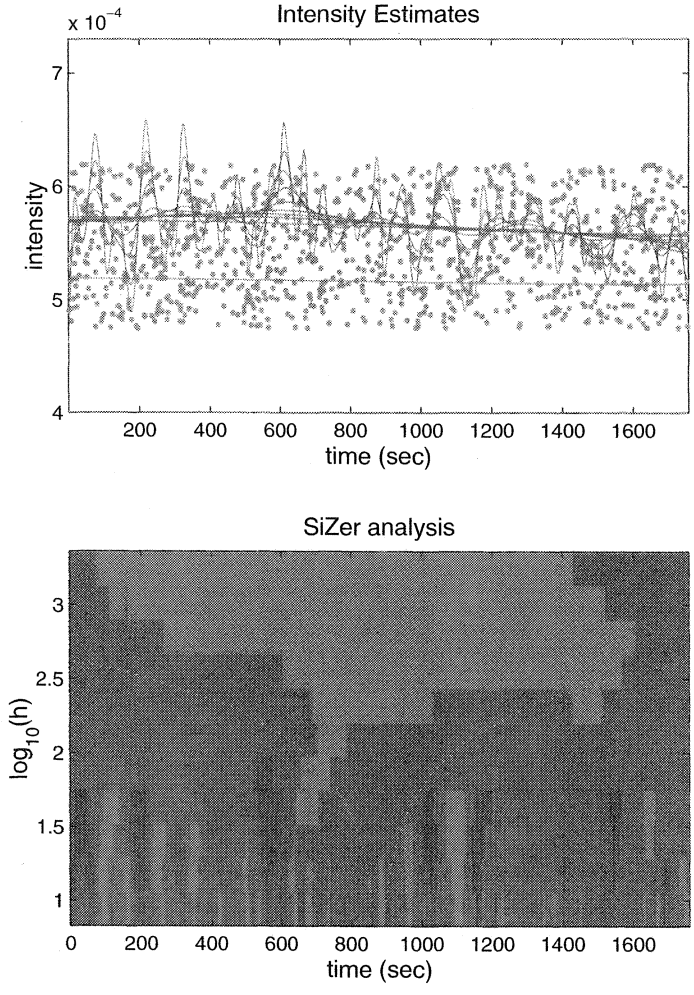

FIGURE 6. *SiZer analysis of start times simulated by Weibull(0.45) renewal model. Shows features similar to real data, but with some differences.*

If one is willing to ignore this additional small scale burstiness it might be tempting to use this as a simulation model for IP flow start times. In this case one should check how well the Weibull (0.45) inter-arrival times approximate the true inter-arrival times. The Q-Q plot, for comparing the original data with the theoretical distribution of Weibull(0.45) is shown in Figure 7.

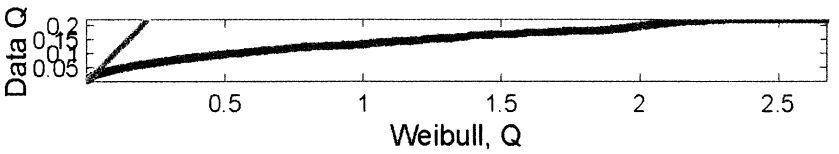

FIGURE 7. *Q-Q analysis of start times simulated by Weibull(0.45) renewal model. Shows this model clearly inappropriate.*

Figure 7 shows that the Weibull (0.45) is an extremely poor fit of the inter-arrival distribution on the data shown in Figure 1. The conclusion of this section is that independence of the inter-arrival distribution is an unworkable assumption.

4. A CLUSTERED POISSON MODEL

There are many possibilities for generation of point processes with dependent inter-arrival times. A sensible approach is to start with simple processes, and to at least initially give most weight to models with substantial physical interpretation. Gennady Samorodnitsky suggested the simple and intuitive Clustered Poisson model as satisfying these criteria. This model was also proposed and studied by Santiee, Nuzman, Sweldens, and Weiss (2001).

The Clustered Poisson model starts with an underlying homogeneous Poisson Process. At each of those event times, a random number of additional "nearby" points is generated and the combined set of points are the events of the full process. This process makes physical sense for many applications that communicate using IP, such as the world-wide web. Web browsers usually make a first request to download the source code of a given web page. If a page has embedded objects (such as graphics, banner ads or internal frames), the web browser opens a new connection for each object download (this behavior corresponds to the non-persistent version of HTTP, which accounted for 85% of all the HTTP flows in the data set. See Smith, Hernandez-Campos, Jeffay and Ott (2001) for more details.) Studies of web traffic, such as Mah (1997), Barford and Crovella (1998) and Smith, Hernandez-Campos, Jeffay and Ott (2001), show that most pages do have embedded objects, and the size of these objects is small. As a consequence, accurate modeling of this flow arrival process is critical.

As a simple first attempt, clusters were simulated according to a Poisson distribution, with mean parameter λ. Because of the above idea of web pages calling for additional flows, cluster points always appear later in time than the initial Poisson point. A very simple distribution for the cluster points is the right triangular density, supported on the interval $[0, \tau]$, with peak at 0. This reflects the idea that cluster points are somewhat more likely to be closer to the original Poisson point, and have a compactly supported distribution (perhaps related to the original flows round trip time). All of the above choices should be regarded as crude first approximations. It would be very interesting to update these assumptions, particularly based upon the study of the underlying processes at work with IP flows.

These assumptions reduce the model choice to only the choice of λ and τ, as well as the Poisson intensity of the original process. For each λ, the latter was chosen to give a similar number of points, over nearly the same time interval as the original data in Figure 1. Choice of the parameters λ and τ, was then done with a trial and error process, with the goal of emulating the SiZer performance of Figure 2. Some of the intermediate steps are interesting, but are not shown here to save space. These can be found in the files UNC2000FlowSimClustPois11s20.ps, UNC2000FlowSimClustPois12s20.ps, UNC2000FlowSimClustPois13s20.ps and UNC2000FlowSimClustPois14s20.ps in the above web directory, but are not shown here to save space.

The best simulated result, using the parameters $\lambda = 16$ and $\tau = 20$, is shown in Figure 8.

The wiggles in the top panel of Figure 8 have a visually similar random structure to those in the top panel of Figure 2. The SiZer map also has a somewhat similar structure for the medium and also small scales (in the sense of a similar distribution). The large scale structure is different, but again this seems to be within the range of "reasonable distributional differences".

Another type of confirmation comes from a Q-Q analysis, as in Figure 4. This showed a reasonable fit of the Weibull distribution, except for some boundary distortion. It is not shown here to save space, but is available in the file `UNC2000FlowSimClustPois14QQlog.ps` in the above web directory.

Hence the Clustered Poisson model is recommended for further work as a candidate model for the simulation of IP flow start times. It is expected that a deeper analysis of

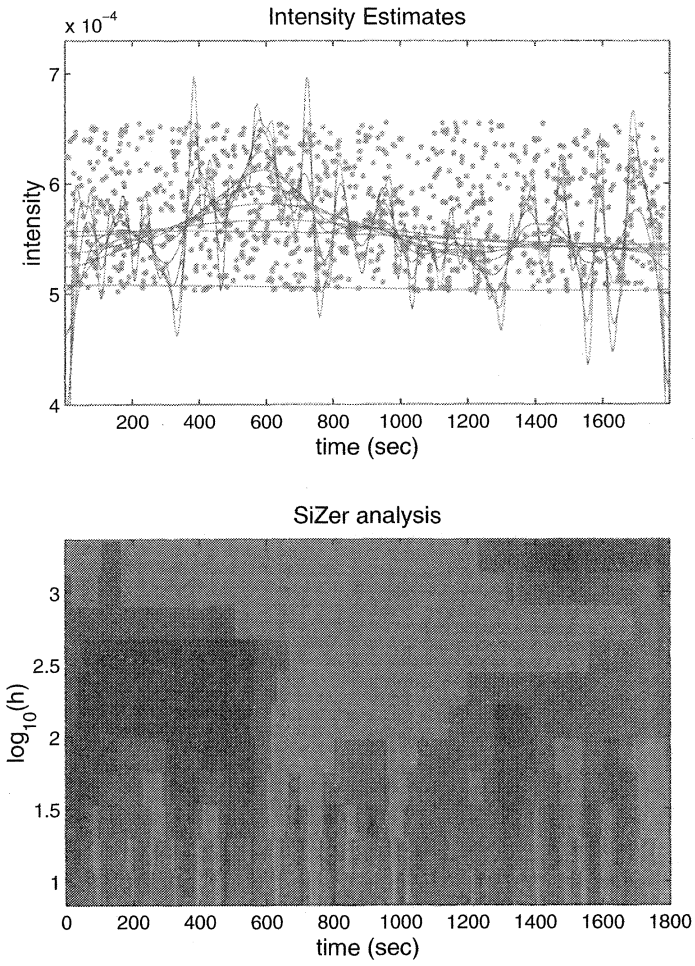

FIGURE 8. *SiZer analysis of start times simulated by Clustered Poisson model. Suggests similar structure to real data, analysed in Figure 2.*

the cluster number parameter λ, and the cluster distribution, will reveal new insights about the working of Internet traffic.

5. ACKNOWLEDGEMENT

The research of J. S. Marron was supported by Cornell University's College of Engineering Mary Upson Fund and NSF Grant DMS-9971649. Most of the data analysis in this paper was done in the stimulating environment of the course OR 778, taught

by the first author in the Fall of 2001 at the Cornell University School of Operations Research and Industrial Engineering.

REFERENCES

[1] Barford, P. and Crovella, M. (1998) Generating representative web workloads for network and server performance evaluation. *Proceedings of the ACM SIGMETRICS Conference, Madison, WI, July 1998.*

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

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

[4] 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.

[5] Cao, J., Cleveland, W. S., Lin, D. and Sun, D. X. (2001) On the nonstationarity of internet traffic, *Proceedings of ACM SIGMETRICS '01*, 102-112.

[6] 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.

[7] Danzig, P. B., Jamin, S., Caceres, R., Mitzel, D. and Estrin, D. (1992) An Empirical Workload Model for Driving Wide-area TCP/IP Network Simulations, *Internetworking: Research and Experience*, 3, 1–26.

[8] Feldmann, A. (2000) Characteristics of TCP connection arrivals, *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, eds., John Wiley, New York.

[9] Garrett, M. W. and Willinger, W. (1994). Analysis, Modeling and Generation of Self-Similar Video Traffic, *Proceeding of the ACM Sigcom '94, London, UK*, 269-280.

[10] Kulkarni, V. G., Marron, J. S. and Smith, F. D. (2000) A Cascaded On-Off Model for TCP Connection Traces, unpublished manuscript.

[11] 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.

[12] Mah, B. A. (1997) An empirical model of HTTP network traffic, *Proceedings of the IEEE International Conference on Computer Communication, Kobe, Japan.*

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

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

[15] Postel, J. (1981) Internet Protocol, *Internet Request for Comments (RFC)*, no. 791, September 1981.

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

[17] Santiee, I., Nuzman, C., Sweldens, W. and Weiss, A. (2001) A compound model for TCP connection arrivals: empirical study with a general framework, to appear in the *Proceedings of ITC 2001*, Brazil, Sep 2001.

[18] Scott, D. W. (1982) *Multivariate Density Estimation Theory, Practice and Visualization*, Wiley, New York.

[19] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[20] Smith, F. D., Hernandez-Campos, F., Jeffay, K. and Ott, D. (2001) What TCP/IP Protocol Headers Can Tell Us About The Web, *Proceedings of the ACM SIGMETRICS/Performance.*

J. S. MARRON
SCHOOL OF OPERATIONS RESEARCH AND INDUSTRIAL ENGINEERING
CORNELL UNIVERSITY ITHACA
NEW YORK 14853
AND DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA CHAPEL HILL
NC 27599-3260

marron@email.unc.edu

FELIX HERNANDEZ-CAMPOS
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-3175

fhernand@cs.unc.edu

F. D. SMITH
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-3175

smithfd@cs.unc.edu