Long-Range Dependence in a Changing Internet Traffic Mix

Cheolwoo Park
Statistical and Applied Mathematical Sciences Institute, RTP, NC

Felix Hernandez-Campos
Department of Computer Science, University of North Carolina at Chapel Hill

J. S. Marron
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

F. Donelson Smith*
Department of Computer Science, University of North Carolina at Chapel Hill

** corresponding author smithfd@cs.unc.edu*

Abstract

This paper provides a deep analysis of long-range dependence in a continually evolving Internet traffic mix by employing a number of recently developed statistical methods. Our study considers time-of-day, day-of-week, and cross-year variations in the traffic on an Internet link. Surprisingly large and consistent differences in the packet-count time series were observed between data from 2002 and 2003. A careful examination, based on stratifying the data according to protocol, revealed that the large difference was driven by a single UDP application that was not present in 2002. Another result was that the observed large differences between the two years showed up only in packet-count time series, and not in byte counts (while conventional wisdom suggests that these should be similar). We also found and analyzed several of the time series that exhibited more "bursty" characteristics than could be modeled as Fractional Gaussian Noise. The paper also shows how modern statistical tools can be used to study long-range dependence and non-stationarity in Internet traffic data.

1. Introduction

The seminal papers reporting empirical evidence for long-range dependence and self-similarity in network traffic first appeared around ten years ago [20]. Their findings were a "disruptive" event in networking research. They discovered a remarkable characteristic of Internet traffic – its *high variability across a wide range of time scales, and how that variability changes as the scale increases*. If we plot the number of packets or bytes that arrive at a network link, say every 10 milliseconds, we observe a highly variable process. Interestingly, if we plot these arrivals at coarser scales, say every 0.1 second, every second, or every 10 seconds, etc., we obtain a rather unexpected result. Instead of smoother and smoother arrival counts as we would expect, we always observe a process that is almost as variable as the one observed at the finer scales. This property of the variance in packet or byte arrivals in Internet traffic, which is known as *self-similarity* or *scale-invariance*, holds true for scales from a few hundred milliseconds up to hundreds of seconds. Quantitatively, the decay in the variance of packet or byte arrival counts in fixed intervals of time for such self-similar traffic is proportional to m^{2H-2}. Here $m \geq 1$ represents the scale of time aggregation of counts, and H is known as the *Hurst* parameter. For a time series of counts generated by a Poisson process (not self-similar), $H=0.5$, while $H \in (0.5,1)$ for a stationary, self-similar process. Values of H > 1 indicate non-stationarity. The closer the value of the Hurst parameter is to 1, the more slowly the variance decays as scale (m) increases, and the traffic is said to be more *bursty*. The slow decay of variance in arrival counts as scale increases in self-similar traffic is in sharp contrast to the mathematical framework provided by Poisson modeling in which the variance of the arrivals process decays as the square root of the scale (see [20], [26]).

Self-similarity also manifests itself as *long-range dependence* (or *long memory*) in the time series of arrivals. This means that there are non-negligible correlations between the arrival counts in time intervals that are far apart. More formally, the autocorrelation function, $\rho(k)$, of long-range dependent time series decays in proportion to $k^{-\beta}$ as the lag k (the distance between elements in the series) tends to infinity, where $0<\beta<1$. The Hurst parameter is related to β via $H=1-\beta/2$, so the closer the value of the Hurst parameter is to 1, the more slowly the autocorrelation function decays. In contrast, Poisson models are short-range dependent, *i.e.*, their autocorrelation decays exponentially as the lag increases. The implied “failure of Poisson modeling” [26] for Internet traffic spawned an active field of research in analysis of network traffic. Some of the research closely related to this paper is reviewed in section 2.

One of the major strengths of the early studies was that they were based on a significant number of high-quality network traces (high quality in the sense that they captured hours or days of operation on production networks and were recorded with reasonably accurate and precise timestamps for each packet). In recent years, however, there have been only a few studies that examined network traffic in the modern Internet using empirical data comparable in quantity and quality to the earlier studies (a few exceptions, notably from Sprint Labs, are described in section 2). There are several reasons for this decline in studies based on empirical data. One is that network links have dramatically increased in speed from the 10 Mbps Ethernets monitored for the early studies to the 1000 Mbps Ethernets and 2500 Mbps (OC-48) or faster technologies commonly deployed in today’s access and backbone network links. Capturing traces, even when such links are lightly loaded and the traces include only packet headers, is costly in terms of the required monitoring equipment and disk storage. Another important issue is that network service providers regard as proprietary or trade secret any empirical data that would reveal information about the operational characteristics of their networks. Similarly, all service providers, including those that are part of universities or other research organizations, have to be concerned about privacy issues. Because of these factors, there are relatively few suitable network traces publicly available to researchers (the NLANR repository [16] does contain a few such traces that have been used in some of the studies reviewed in section 2).

In this paper we present the results from analysis of an extensive set of two-hour traces collected in 2002 and 2003 from a high-speed (1000 Mbps) access link connecting the campus of the University of North Carolina at Chapel Hill to its Internet service provider (ISP). The traces were started at the same four times each day for a week-long period. In aggregate, these traces represent 56 hours of network activity during the second week of April in 2002 and 56 hours during the same week of April in 2003. The 2002 traces include information about 5 billion packets that carried 1.6 Terabytes of data while the 2003 traces represent almost 10 billion packets and 2.9 Terabytes (the load on the Internet link almost doubled in one year). The details of the trace acquisition methods and summary statistics of the traces are presented in section 3 and Appendix A.

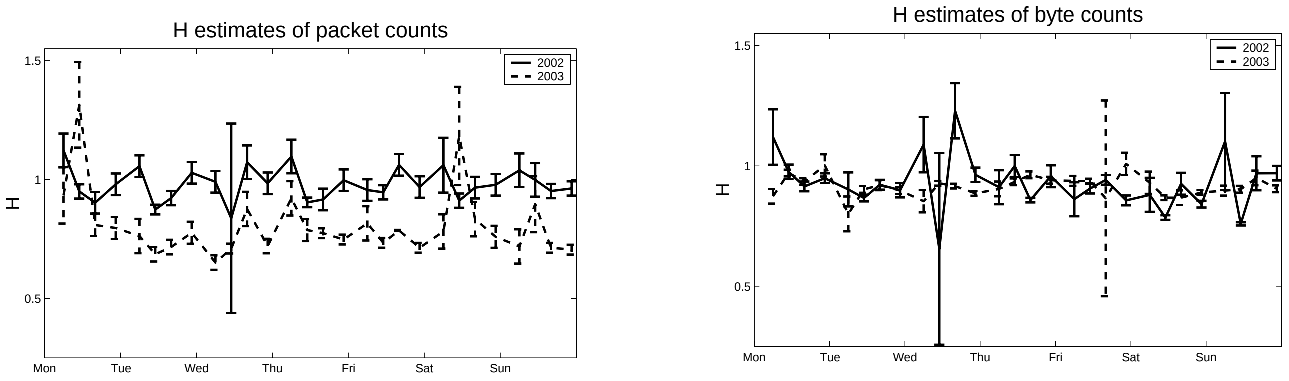


Figure 1. Hurst parameter estimates and confidence intervals for packet and byte counts by day of week and time of trace start. Clearly the H estimates for packet counts were significantly lower in 2003 while the H estimates of byte counts showed little differences between the two years.

Figure 1 provides a visual summary of the estimated Hurst parameter (H) and 95% confidence intervals (CI) for the time series derived from each of the traces at the day and time (x- axis) that the trace was obtained. The most surprising observation is that the 2003 packet-count time series generally had significantly lower estimated H values than in 2002. The H estimates for the byte counts were, however, virtually identical between the two years which is also unexpected because previous studies have shown packet and byte counts to be closely related. Investigating the reasons for these surprising results is a major theme of this paper.

We also see in Figure 1 that a number of traces exhibited large H values (> 1.0) or had very wide confidence intervals for the H estimates. Hurst parameter estimates where H > 1 are problematic in several ways. While such stochastic processes exist, they are non-stationary and the estimation methods used here may be unreliable. We investigate this point carefully using the *Dependent SiZer* analysis and visualization methods (described in section 5). Several types of non-stationarity were found using this method. Some of the non-stationarities were simple time series trends, such as a linear increase or decrease. This type of trend is easily filtered out by the wavelet-based tools we used (described in section 4). Other types of non-stationarity were also encountered and these were shown by Dependent SiZer to be caused by phenomena such as occasional local bursts whose magnitude is inconsistent with conventional long-range-dependent models such as Fractional Gaussian Noise (FGN). A natural question is: does it still make sense to estimate H in the presence of such phenomena? We have found that it does. One reason is that large H estimates, or wide confidence intervals for the H estimate, are good indicators that problematic features exist in the data. Another reason is that the Dependent SiZer analysis shows that even when such phenomena occur, they tend to be quite local in time while standard models such as FGN apply over most time intervals. Thus H estimation still remains worthwhile even for time series in which its value or confidence interval indicates non-stationarity.

The statistical tools used for this paper include a wavelet-based spectrum analysis and Hurst parameter estimation (as a simple quantification of long-range dependence), Dependent SiZer analysis for specific investigations of structure beyond that indicated by the Hurst parameter, and a cross-trace summary plot of SiZer structure for use in the simultaneous analysis of a large data set. We used the wavelet-based analysis tools that were developed by Abry and Veitch [1]. SiZer was invented by Chaudhuri and Marron [7], and the dependent-data version that is used in section 5 was developed by Park, Marron and Rondonotti [25].

Our primary findings were:

- Hurst parameter estimates for the packet counts in 10 millisecond intervals decreased significantly between 2002 and 2003 while the Hurst parameter estimates for byte counts remained about the same.
- A single peer-to-peer file sharing application (*Blubster*) that suddenly became popular between 2002 and 2003 had a strong influence on the Hurst parameter for packet count time series and caused the differences in H between the two years. This happened because the application sends a large number of small packets with high frequency that, in turn, produced high variance at certain time scales.
- Moderate to strong long-range dependence (Hurst parameter estimate in the range [0.65, 0.99]) existed in the vast majority of the time series.
- In several of the 56 traces studied, we found evidence of highly bursty traffic in packet or byte counts that could not be modeled as Fractional Gaussian Noise.
- The long-range dependence, as measured by the Hurst parameter estimate, was not influenced by the day-of-week or time-of-day. This implied that changes in the *active* user population (and, by implication, types of applications used) did not influence long-range dependence.
- The long-range dependence as measured by the Hurst parameter estimate was not influenced by the traffic load (link utilization) or the number of active TCP connections. In particular we found no indication of declining long-range dependence as more active TCP connections are aggregated.

We have also demonstrated the value of employing a suite of statistical tools that includes wavelet analysis and Dependent SiZer to fully explore the nature of long-range dependence in Internet traffic.

The remainder of this paper is organized as follows. Section 2 briefly reviews some additional research that is related to the themes of this paper. Section 3 describes details of the tracing methods and presents summary statistics of all the trace data. Section 4 gives a brief summary of the wavelet-based methods for analysis. It also presents summary results for aggregated packet- and byte-count time series for all the traces. Section 5 focuses on the use of Dependent Sizer to analyze complex structure in several packet- and byte-count time series. In section 6 we present results for protocol-dependent subsets focusing on the differences between the time series of all packets and bytes compared with just TCP or UDP packets and bytes. We also show how a single peer-to-peer application influences the Hurst parameter. Section 7 summarizes our findings.

2. Other Related Research

The early studies that found empirical evidence for self-similarity and long-range dependence spawned an active field of research in analysis of network traffic. Various issues were pursued including those seeking evidence for pervasive long-range dependent properties (e.g., [26]), finding physical explanations for the causes (e.g., [8], [22] and [30]), investigating the implications for network and protocol designs or performance analysis (e.g., [9], [13] and [23]), and studying scaling behavior at different time scales (e.g., [11]). The book by Park and Willinger [24] (and the references therein) is an excellent resource and guide to the results from this research.

Following the initial studies, attention gradually shifted from studying the well-established scaling properties of traffic at larger time scales (one second and longer) to focus on shorter time scales. Some early investigations attributed the observed complex scaling properties of traffic at sub-second times to *multifractal* scaling. In [11] an empirical validation was given for a claim that networks can be modeled as semi-random cascades that give rise to the observed multifractal scaling behavior. In a follow-up study [12], extensive simulations were used to investigate network-dependent variability as the cause of multifractal scaling at small time scales. Their results pointed to TCP dynamics such as “ACK compression”, especially as influenced by round-trip times, as a major source of burstiness at small time scales.

Even more recently, a different perspective on models for Internet traffic has emerged from empirical studies that used traces from links in the Internet core where traffic from many sources is highly aggregated. These results considered together show that, at least for core links of the Internet, the small-time scale variability of network traffic tends to be more Poisson-like than was expected from the earlier results. In particular, one of the most data-rich studies (from Sprint Labs using extensive high-quality traces from the Sprint Internet backbone) found that traffic variations tended to be un-correlated and have monofractal scaling at small time scales [31]. This study also identified the proportion of *dense* flows (those generating bursts of densely clustered packet arrivals) *vs* sparse flows in the traffic aggregate as the determining factor in the amount of correlation present. Studies from Bell Labs ([4], [5] and [6]) attributed the tendency of packet inter-arrival times to become independent and exponential (Poisson) to increasing numbers of active TCP connections sharing a link. Another recent study of backbone traffic [19] claimed that a stationary Poisson model characterized traffic arrivals up to sub-second time scales and proposed a time-dependent Poisson model to characterize the non-stationarity observed at multi-second time scales. All of these studies have, however, confirmed the invariant property of long-range dependence in packet or byte counts at time scales above one second and these longer scales are the focus of this paper.

In addition to related work in traffic analysis, our research builds on prior work that developed several statistical methods and tools. These have been referenced and discussed in the introduction and in the sections where they are used for our analysis.

3. Traffic Measurements and Data

The two trace collections were obtained by placing a network monitor on the high-speed link connecting the University of North Carolina at Chapel Hill (UNC) campus network to the Internet via its Internet service provider (ISP). All units of the university including administration, academic departments, research institutions, and a medical complex (with a teaching hospital that is the center of a regional health-care network) used a single ISP link for Internet connectivity. The user population is large (over 35,000) and diverse in their interests and how they use the Internet — including, for example, email, instant messaging, student "surfing" (and music downloading), access to research publications and data, business-to-consumer shopping, and business-to-business services (*e.g.*, purchasing) used by the university administration. There are over 50,000 on-campus computers (the vast majority of which are Intel architecture PCs running some variant of Microsoft Windows). For several years the university has required all incoming freshmen to have a personal computer, typically a laptop with wired or wireless network interfaces. There are over 250 wireless access points operating at 11 Mbps throughout the campus and a significant fraction of Internet traffic transits the wireless network. Wired connections are 100 Mbps and are pervasive in all buildings including the dormitories. The core campus network that links buildings and also provides access to the Internet is a switched 1 Gbps Ethernet.

We used a network monitor to capture traces of the TCP/IP headers on all packets entering the campus network from the ISP (we call these packets "inbound" traffic in this paper). All the traffic entering the campus from the Internet traversed a single full-duplex 1 Gbps Ethernet link from the ISP edge router to the campus aggregation switch. In this configuration, both "public" Internet and Internet 2 traffic were co-mingled on the one Ethernet link and the only traffic on this link was traffic from the ISP edge router. We placed a monitor on this fiber link by passively inserting a "splitter" to divert some light to the receive port of a Gigabit Ethernet network interface card (NIC) set in "promiscuous" mode. The *tcpdump* program was run on the monitoring NIC to collect a trace of TCP/IP packet headers. The trace entry for each packet included a timestamp (with 1 microsecond resolution and approximately 1 millisecond accuracy), the length of the Ethernet frame, and the complete TCP/IP header (which includes the IP length field).

For the 2002 traces, the monitor NIC was hosted in an Intel-architecture server-class machine configured with a 1.0 GHz processor, 1 GB of memory, 64 bit-66 MHz PCI busses, six 10,000 RPM 31GB disks and running FreeBSD version 4.1. Buffer space of 32KB was allocated to the *bpf* device used by *tcpdump* to buffer transient overloads in packet arrivals. The *tcpdump* program reports statistics on the number of packets dropped by *bpf.* We found that in only 8 of the 28 traces was the drop rate 0.01% or less (1 packet per 10,000) and the drop rate was as high as 0.16% in one trace. For the 2003 traces we upgraded the monitor machine to a 1.8 GHz processor, 64-bit 100 MHz busses and five 15,000 RPM 35 GB disks. The *bpf* buffer was set to 16 MB and the *tcpdump* program was modified to use block writes of 512 bytes. As a result no packets were dropped in any of the 2003 traces in spite of an almost two-fold increase in the load on the link.

The traces were collected during four two-hour tracing intervals each day for seven consecutive days of the week (28 traces per collection period). Both the 2002 and 2003 collection periods were during the second week in April, a particularly high traffic period on our campus coming just a few weeks before the semester ends. Collecting traces during the same week in April of both years allowed us to compare results from traces gathered one year apart. The two-hour intervals were 5:00-7:00 AM, 10:00-12:00 noon, 3:00-5:00 PM, and 9:30-11:30 PM. These periods were chosen somewhat arbitrarily to produce two traces during high traffic periods of the normal business day, and two during non-business hours when traffic volumes were the lowest (5:00-7:00 AM) or "recreational" usage was likely to be high (9:30-11:30 PM).

Appendix A provides summary information about the complete set of 56 traces including the total packets and bytes with a breakdown by transport protocol, TCP *vs* UDP. Also included are the average link utilization (relative to an idealized 1 Gbps transmission speed) and an estimate of the median number of TCP connections per minute over the tracing interval[1]. We found that the highest average link utilization over an entire trace almost doubled (from 10% to 18%) between 2002 and 2003 reflecting the growth in Internet usage on the campus. Likewise, the number of active TCP connections per minute also increased substantially in 2003 compared to 2002. Perhaps the most striking observation was the growth in UDP over the period of a single year from around 5% of packets and 7% of bytes to about 25% of packets and 14% of bytes. Clearly, the growth came from UDP applications that generated a smaller average packet size than had previously been present. Furthermore, data collected from this same link for the previous 5 years had shown that the average usage of UDP in both packets and bytes had remained nearly constant at levels similar to the 2002 data. We explore the causes and implications of this growth in UDP in section 6.

4. Wavelet-based Estimation of Hurst Parameters

The concepts and definitions for self-similarity and long-range dependence given in section 1 assume that the time series of arrivals is *second-order stationary* (a.k.a. *weakly stationary*). Loosely speaking, this means that the variance of the time series (and more generally, its covariance structure) does not change over time, and that the mean is constant (so the time series can always be transformed into a zero-mean stochastic process by simply subtracting the mean). The obvious question this raises is whether Internet traffic is stationary. This is certainly not the case at the scales in which the time-of-day effects are important (traffic sharply drops at night), so Internet traffic is usually characterized as self-similar and long-range dependent only for those scales between a few hundred milliseconds and a few thousand seconds. Furthermore, we often find trends and other effects even at these time scales. For example, the UNC link showed an increase in traffic intensity during morning hours as more and more people become active Internet users. However, it is still useful to study time series using the self-similarity and long-range dependence framework, and this is possible using methods that are *robust* to non-stationarities in the data (*i.e.*, methods that first remove trends and other effects from the data to obtain a second-order stationary time

[1] Estimating valid TCP connections per minute is complicated by the presence of SYN-flooding and port-scanning attacks. We estimated the number of TCP connections per minute by computing the median of min(count of TCP segments with SYN flag, count of TCP connections with FIN or RST flag) for each 1 minute interval. SYN-flooding and port-scanning attacks usually have abnormally high counts of SYN or RST segments in one direction but not both.

series). In some cases, non-stationarities are so complex, that conventional models fail to accommodate them, which can result in estimates of the Hurst parameter greater than or equal to 1. We examine some of these cases in section 5.

The wavelet-based tools for analysis of time series are important because they have been shown to provide a better estimator (and confidence intervals) than other approaches for the Hurst parameter [14]. These methods also are robust in the presence of certain non-stationary behavior (notably linear trends) in the time series. This is particularly important in working with traces having durations of two hours because we observed linear trends caused by diurnal effects in many of them (*e.g.*, a ramp-up of network activity between 10:00 AM and noon).

For the results in this paper we used the Abry and Veitch methods (and MATLAB software) [28] to study the wavelet spectrum of the time series of packet and byte counts in 10 millisecond intervals. The output of this software is a *logscale diagram*, *i.e.*, a *wavelet spectrum*, which provides a visualization of the scale-dependent variability in the data. Briefly, the logscale diagram plots the $\log_2$ of the (estimated) variance of the Daubechies wavelet coefficients at a scale value against $j = \log_2(\text{scale})$ where j is often called the *octave*. For processes that are long-range dependent, the logscale diagram will exhibit a region in which there is an approximately linear relationship with positive slope at the right (coarser scale) side. An estimate of the Hurst parameter, H, along with confidence intervals on the estimate can be obtained from the slope of this line (H=(slope+1)/2). The choice of scales, j_1 and j_2, that are endpoints for this line determine the scaling region for the Hurst parameter. The Abry and Veitch methods [29] (and MATLAB software) we used automatically selects the j_1 and j_2 values for each H estimate. For Internet traffic, the region where this linear scale relationship begins is usually above one second (we note that a different scaling relationship below one second has been observed by others in Internet traffic – we do not consider this “bi-scaling” effect in this paper). For the vast majority of traces (both packets and bytes), j_1 was automatically set to 8 or 9 and j_2 was set to 16. A few of the traces have a significantly different j_1 - j_2 range (e.g., 11-16) which causes larger H estimates and/or wider confidence intervals. This is an indication of non-stationarities and we explore them using dependent SiZer in section 5. Since the scale effectively sets the time-aggregation level at which the wavelet analysis is applied, there is a direct relationship between scale and time intervals. The wavelet coefficients were computed for our two-hour traces typically using scales from 2^8 to 2^{16} (from 256 to about 65,500 of our 10 millisecond intervals or between 2.5 and

655 seconds). For a more complete summary of long-range dependence and wavelet analysis see also Chapter 2 of [24], [1], [2] and [27].

4.1 Summary Results

Each of the traces was processed to produce two time series, one for packet counts and one for byte counts, both in 10 millisecond intervals (a 10 millisecond interval is within the accuracy of our monitoring infrastructure). For each of the resulting time series, we used the wavelet analysis tools to determine if the time series exhibited long-range dependence and to compute the Hurst parameter estimate (H) along with 95% confidence intervals (CI). We set the parameter (N) that controls the number of vanishing moments of the wavelet to 3 as recommended in [28] in order to avoid potential bias caused by the trends we observed in most of the two-hour traces. Using three vanishing moments of the wavelet can remove polynomial trends up to degree 2.   In the vast majority of cases the estimated H values and their confidence intervals fall in the range [0.65, 0.99] indicating moderate to strong long-range dependence. There are a few values of the estimated Hurst parameter in the range [1.0, 1.1] but where the CI range includes values < 1.0. We found a few others that represent more extreme cases in the sense that either the estimated H value and its CI range was > 1.0 (indicating non-stationarity that the wavelet tools could not eliminate) or the confidence interval was extremely wide. We look at some of these in more detail in section 5.1

Figure 1 (in section 1) provides a visual summary of the estimated Hurst parameter (H) and confidence intervals (CI) for each of the time series at the day and time (x- axis) that the trace was obtained. There are several interesting features, notably that the 2003 packet-count time series generally had significantly lower estimated H values than in 2002. The H estimates for the byte counts were, however, virtually identical between the two years. The reasons for these differences in H estimates are further explored in section 6. There did not appear to be any consistent relationships between the day of week or the time of day and the estimated value for H. This implied that changes in the makeup of the *active* user population at different days and times (*i.e.*, changes in the proportions of faculty, staff and student users) and, by implication the types of applications used, did not influence long-range dependence. Further, night hours were expected to have more “recreational” uses of the network but this appeared to have no effect. Figure 2 shows the estimated value of H for packet and byte counts as a function of the average link utilization during the trace interval and Figure 3 shows them as a function of the average number of TCP connections per minute. Clearly both link utilization and active TCP connections increased substantially between 2002 and 2003 but there did not appear to be any evidence of the Hurst estimate being influenced by these changes. Overall, we found no evidence that link utilizations or the aggregations of large numbers of TCP connections had any effect on long-range dependence as measured by the Hurst parameter.

`

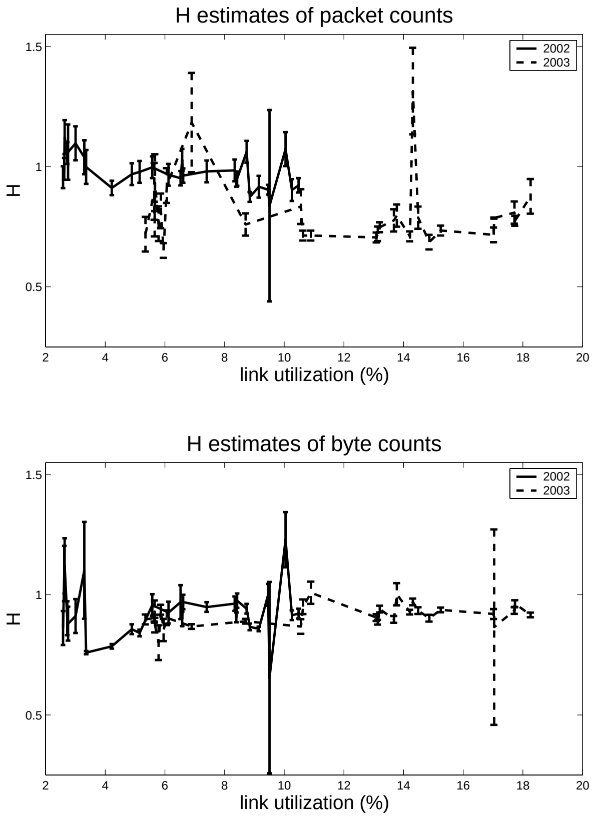


Figure 2. Hurst estimates and CI for packet and byte counts by link utilizations. There was no effect on H estimates of different link utilizations.

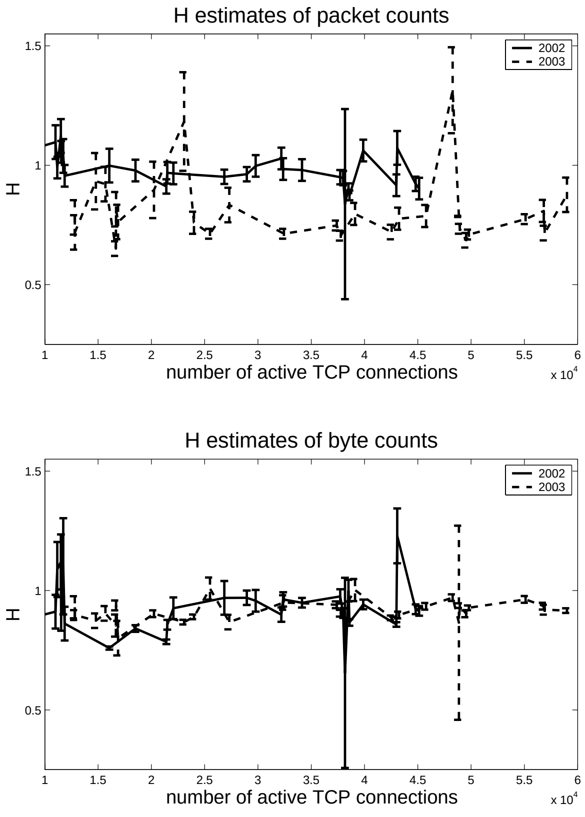


Figure 3. Hurst estimates and CI for packet and byte counts by active TCP connections. There was no effect on H estimates of different TCP connection loads.

5. Dependent Sizer Analysis of Time Series

We carefully examined the six time series summarized in Table I for additional structure not consistent with long-range dependence using the SiZer approach to exploratory data analysis. SiZer provides a useful method for finding *statistically significant* local trends in time series and is especially useful for finding important underlying structure in time series with complex structure. SiZer is based on local linear *smooths* of the data, shown as curves corresponding to different window widths in the top panel of Figure 4 (a random sample of time series values is also shown using dots). These curves are very good at revealing potential local trends in the time series, and provide a visualization of structure in the time series at different *scales*, *i.e*. at different window widths used for smoothing. See [10] for an introduction to local linear smoothing. Two important issues are: which of these curves (*i.e.,* scales) is the *right* one, and which of the many visible trends (at a variety of different scales) are statistically significant (thus representing important underlying structure) as opposed to reflecting *natural variation*? (*i.e.,* unimportant artifacts of the underlying noise process).

Trace	Type	H and CI
Wednesday 10:00 AM, 2002	packets	0.84 [0.44, 1.24]
Wednesday 10:00 AM, 2002	bytes	0.65 [0.26, 1.05]
Wednesday 3:00 PM, 2002	bytes	1.23 [1.11, 1.34]
Monday 10:00 AM, 2003	packets	1.31 [1.13, 1.49]
Friday 3:00 PM, 2003	bytes	0.87 [0.46, 1.27]
Saturday 10:00 AM, 2003	packets	1.18 [0.98, 1.39]

Table I. Traces with H estimates or CI ranges that represent extreme examples

Choice of the window width in curve smoothing (*i.e*. choice among the overlaid curves) is well known to be a very hard problem in statistics (and has been quite controversial, *e.g*., [18] and [21]). SiZer avoids this problem by taking a scale space approach, where a large family of smooths, indexed by the window width, is used. Essentially, each of these curves represents a different “scale” or “level of resolution” of the data, and all are important, and thus should be used in data analysis. The second part of SiZer (shown in the lower panel of Figure 4) is a graphical device for doing the needed statistical inference, in particular flagging trends in the smooths when they are statistically significant. This is a gray-scale *map*, where the horizontal axis represents time, and thus is the same as the horizontal axis in the top panel. The vertical axis represents scale of resolution, *i.e*., window width of the local linear smooth (more precisely, $\log_{10}(h)$ where h is the standard deviation of the Gaussian window function) with the finest scales at the bottom. At each scale-time location (*i.e*., pixel in the map) statistical inference is done on the slope of the corresponding curve. When there is a significantly positive slope, *i.e*.

there is an *important* upward trend, the dark shade is used. Where the slope is significantly negative (downward trend), the pixel is shaded light. When there is no significant slope, *i.e.*, the variation is what is *naturally expected*, an intermediate gray shade is used. It is important to note that SiZer is inherently a multi-scale method and that the scales represented here overlap with the scales used in the wavelet-based estimates of the Hurst parameter as described in section 4.

An important issue is the definition of *natural variation*. The original SiZer [7] tested against a null hypothesis of white noise (*i.e.*, independent Gaussian). This is clearly inappropriate for Internet traffic, because such time series typically have large amounts of serial dependence. Thus, conventional SiZer maps would show a large amount of "significant structure" (*i.e.*, lots of light and dark regions). SiZer becomes more useful for such time series, when the null hypothesis is changed to a more typical model for network traffic, such as Fractional Gaussian Noise (FGN). This is the idea behind the Dependent SiZer tool [25] used for all the results described here. The Dependent SiZer map in the bottom panel of Figure 4 assumed an underlying FGN with Hurst parameter of H=0.8. This Hurst value is quite reasonable, because it is typical of those shown in Tables II and III in Appendix A considering both 2002 and 2003 results and choosing a single value facilitates comparisons among all the traces. The FGN model also requires a variance parameter, which we estimated from each time series as discussed in [25] and [14].

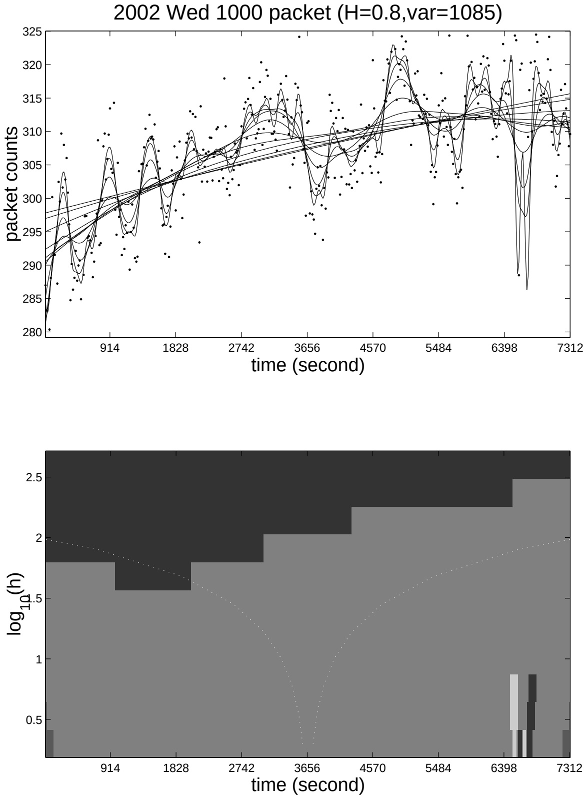


Figure 4. Dependent SiZer analysis of 2002 Wed 10:00 AM packet counts. Significant nonlinear upward trend is atypical of FGN and may have been the cause of the wide confidence intervals for the Hurst parameter estimate.

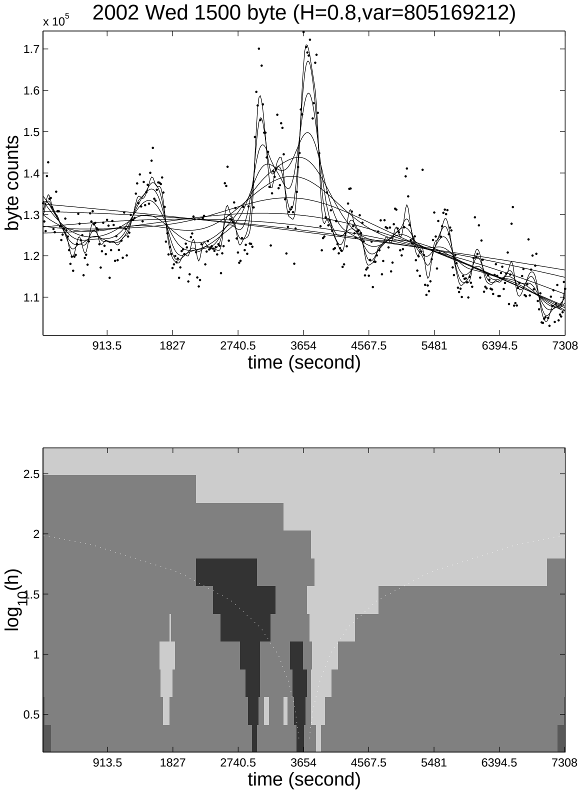


Figure 5. Dependent SiZer analysis of 2002 Wed 3:00 PM byte counts. Spikes were larger than expected for FGN and may have caused the large Hurst parameter estimate.

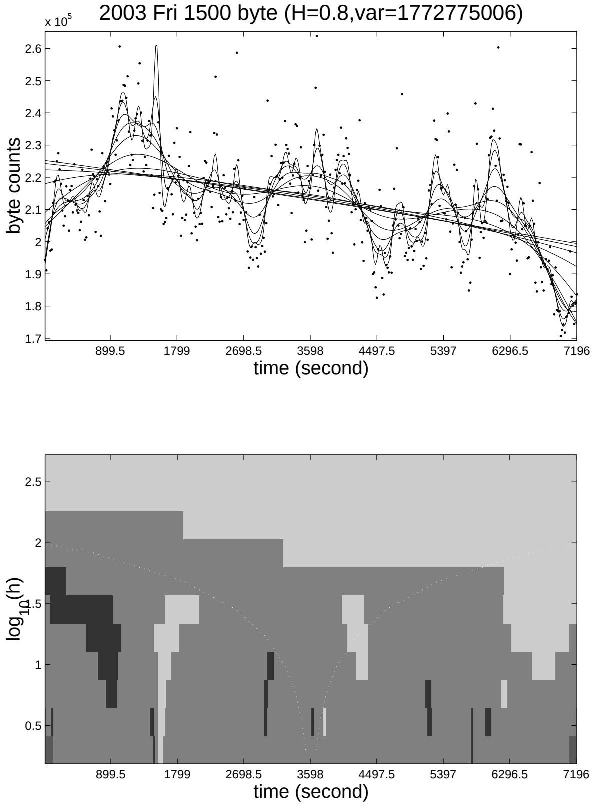
5.1 Analysis of Extreme Cases of H estimates or Confidence Intervals

The results of this analysis applied to the traces in Table I[2] are shown in Figures 4-7. Figure 4 shows a Dependent SiZer analysis of the 2002 Wed 10:00 AM packet-count time series. Recall that this was found to have a very wide confidence interval on the Hurst parameter estimate. The map in the bottom panel shows darker shade near the top, *i.e.*, at the coarser scales. This shows that the upward trend visible in the family of smooths is statistically significant. At first glance it is tempting to view the upward trend as linear, which should have no effect on the Hurst parameter because such effects are filtered out by the wavelet basis used in the estimation process. But a closer look shows that these trends are non-linear and coarse-scale instability may be what is driving the large confidence intervals. There was also some fine-scale behavior flagged as significant in the lower right of the map that corresponds to a substantial valley in the family of smooths. The shades in the map show that both the upward trend and the sharp valley are features that are not typically expected from FGN. However, the other spikes and valleys in the family of smooths correspond to medium gray regions in the map, and thus represent a level of variation that is natural for the FGN model. The significant non-FGN behavior is consistent with the poor behavior of the Hurst parameter estimation process. It is not clear if the overall upward trend or the small sharp valley contributed more to the unusually wide confidence interval. The analysis for 2002 Wed 10:00 AM byte-count time series (not shown) was quite similar. This is not surprising; large aggregations of TCP connections tend to generate very stable distributions of packet sizes so the packet and byte time series tend to be correlated. More discussion about causes of wide confidence intervals is found in the discussion of Figure 8 (below).

Figure 5 shows the Dependent SiZer for the 2002 Wed 3:00 PM byte-count time series. It attracted attention because of the unusually high value of the estimated Hurst parameter. There is less coarse-scale trend, but more shadings in the map (bottom panel) indicating significant medium- and fine-scale features generated by some large spikes seen in the top panel. Again the maps show that these spikes were much larger than would be typical for a FGN model, while the other spikes were consistent with FGN. This suggests that a potential cause of unusually large Hurst parameter estimates is a number of sharp changes that are too large to be consistent with FGN.

Figure 6 shows the analysis for the 2003 Friday 3:00 PM byte-count time series which had an unusually wide confidence interval for the estimated Hurst parameter. The structure is quite different from Figure 4 with a number of smaller peaks and valleys being flagged as statistically significant. This shows that such departures from FGN can also generate a wide confidence interval for the Hurst parameter. Figure 7 shows another case with very large H, which was generated by a much different type of anomaly than seen in Figure 5. The large Hurst parameter estimate seems to have been driven by the substantial non-stationarity of a long increase, followed by a very large drop-off, followed by a nearly constant region.

[2] The 2002 Wed 10:00 AM byte-count analysis is not shown here because it is very similar to 2002 Wed 10:00 AM packet-count analysis. Similarly the



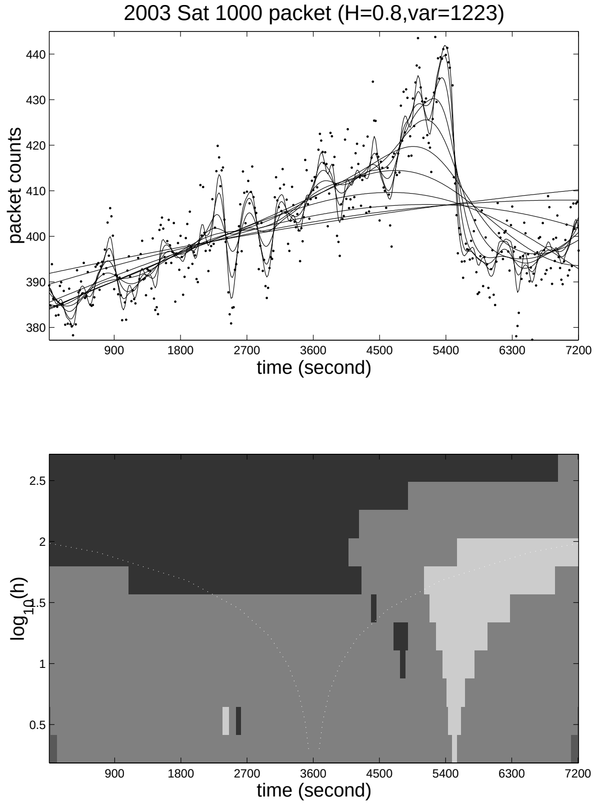


Figure 6. Dependent SiZer analysis of 2003 Fri 3:00 PM byte counts. The many small scale peaks and valleys are statistically significant and may have caused the wide confidence interval.

Figure 7. Dependent SiZer analysis of 2003 Sat 10:00 AM packet counts. Other phenomena, such as a large temporary increase followed by a large drop, were statistically significant and may have caused the large Hurst parameter estimate.

The overall lesson from Figures 4-7 is that a number of different departures from the FGN model can drive the observed large Hurst parameter estimates, and wide confidence intervals. Dependent SiZer provides a clear analysis of which aspects of the data caused this behavior. While most time series were consistent with Fractional Gaussian Noise, our results show that different models were needed in a significant number of cases.

2003 Mon 10:00 AM packet-count analysis is not shown here because the main ideas are similar to 2002 Wed 3:00PM byte-count analysis.

Figure 8 shows the logscale diagrams for all six time series from Table I (confidence intervals for the variance at each scale are shown only for one time series in each plot for readability – the others are very similar). The interesting feature of these logscale diagrams is that each had one “peak” that occurred at octave 12, 14, or 15 (time scales of 40, 160, or 320 seconds). For example, the logscale diagram of the 2002 Wed 10:00 AM packet-count time series has a peak at octave 12, and this roughly matches the duration (about 40 seconds) of the sharp valley in Figure 4. Since the wavelet analysis tools with N=3 (vanishing moments) should be effective in compensating for trends, we conclude that a few intervals of extreme burstiness in the traffic at these time scales lead to the wide confidence intervals for the Hurst parameter.

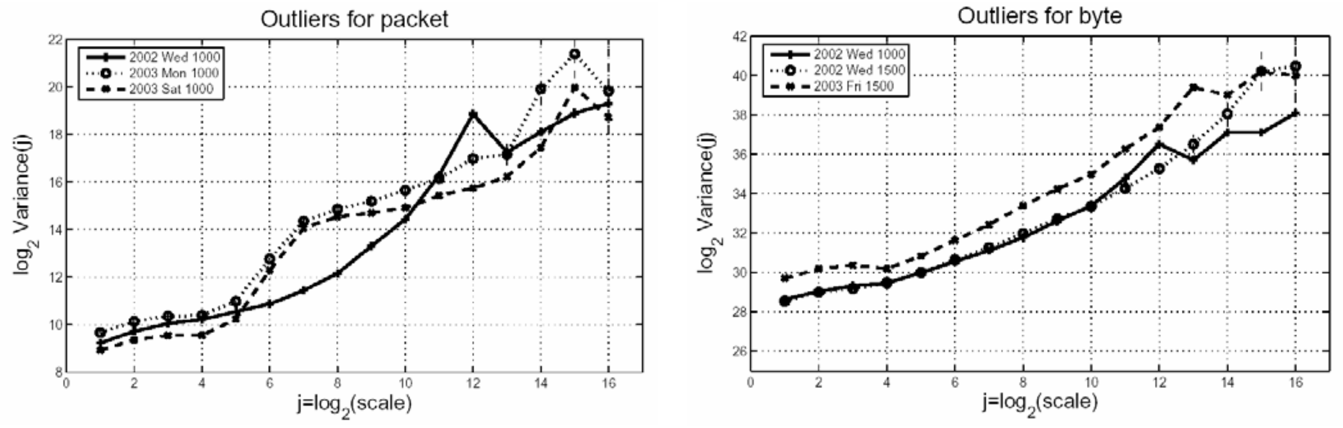


Figure 8. Logscale diagrams for the six time series that represent extreme examples of H estimates or CI width. Peaks at octaves 12, 14, or 15 roughly correspond to significant spikes or valleys in the Dependent SiZer analysis.

5.2 Analysis of Packet Counts for 2002 *vs* 2003

We now consider possible reasons why the H estimates for the packet-count time series were consistently higher for 2002 than for 2003. One possibility for the differences in H estimates for packet counts between the two years is that some differing structural feature of the data was the cause. To investigate these issues we used Dependent SiZer to test each of the time series of packets counts (except those in Table I) for statistically significant differences from a FGN process with H=0.80 (this choice was explained at the beginning of this section). We present here the results for only four cases shown in Figures 9-12: Thursdays at 3:00-5:00 PM (a heavy traffic period) and Saturdays at 9:30-11:30 PM (a light traffic period) for both 2002 and 2003. These four are representative of the results we obtained from examining the complete set.

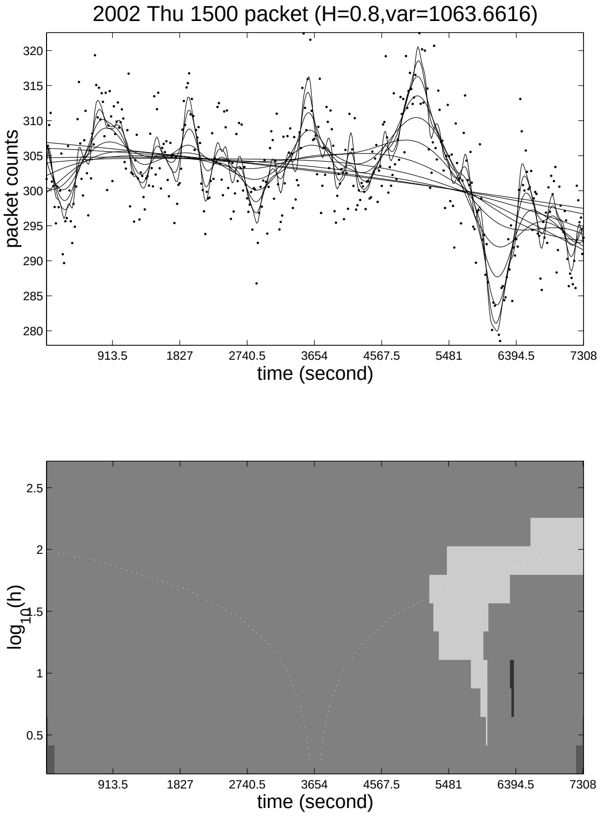


Figure 9. Dependent SiZer analysis for Thursday 3:00 PM, 2002. A FGN model was inadequate to explain the sharp valley near the end of the time series.

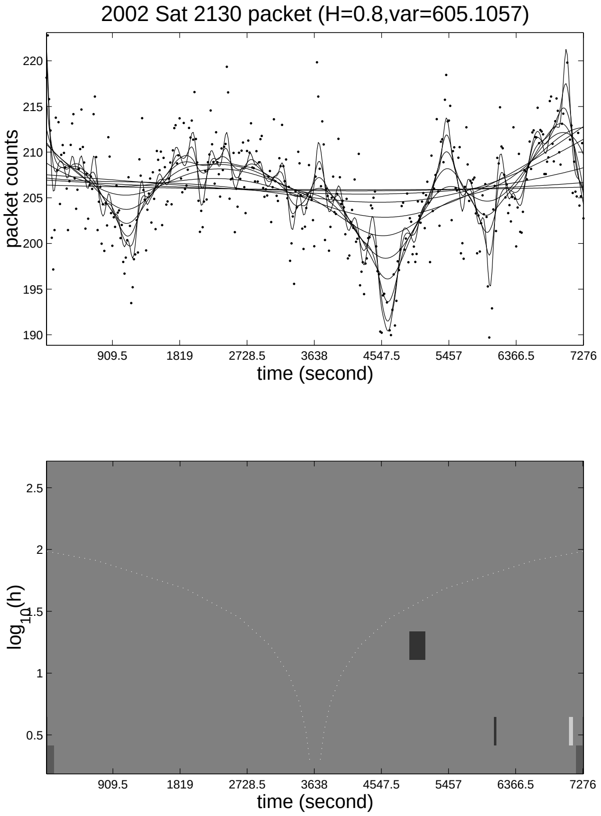


Figure 10. Dependent SiZer analysis for Saturday 9:30 PM, 2002. FGN was essentially appropriate over the entire time series.

In Figure 9, the map shows "natural variation" for the first two thirds of the time span. However the shape decrease about three-quarters of the way through the time interval, around 4:30 PM, is statistically significant, as indicated by the large light region, across a broad range of scales (*i.e.*, the deep valley is a type of variation that is larger than that which is typical of FGN). The smaller dark region shows that there was also a significant increase around 4:40 PM. Figure 10 shows the same analysis for Saturday at 9:30 PM. While the family of smooths in the top panel looks qualitatively similar to that in Figure 9, the map in the bottom panel shows a rather different result. In particular, there is nearly no statistically significant trend or other structure, *i.e.*, FGN is essentially appropriate over this entire time series.

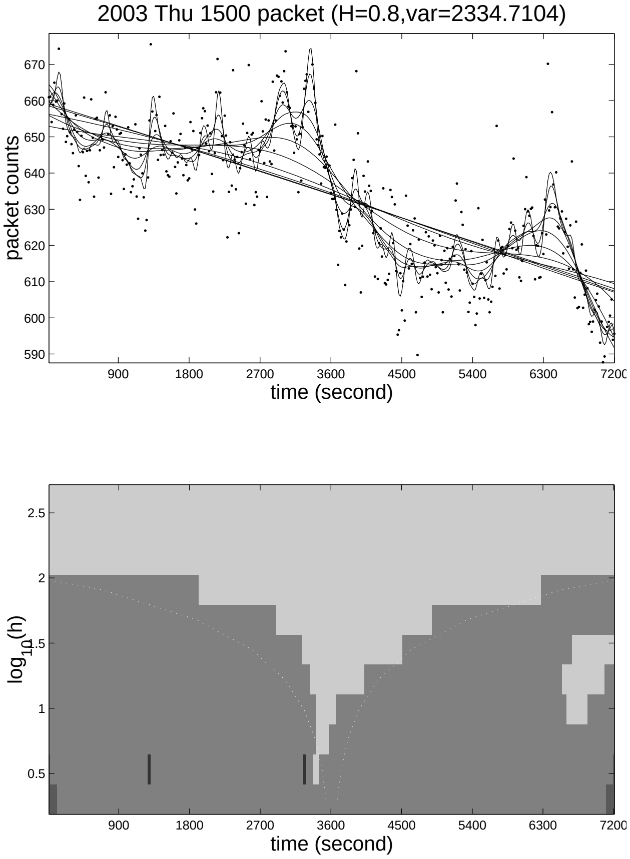


Figure 11. Dependent SiZer analysis for Thursday 3:00 PM, 2003. A statistically significant downward trend is shown.

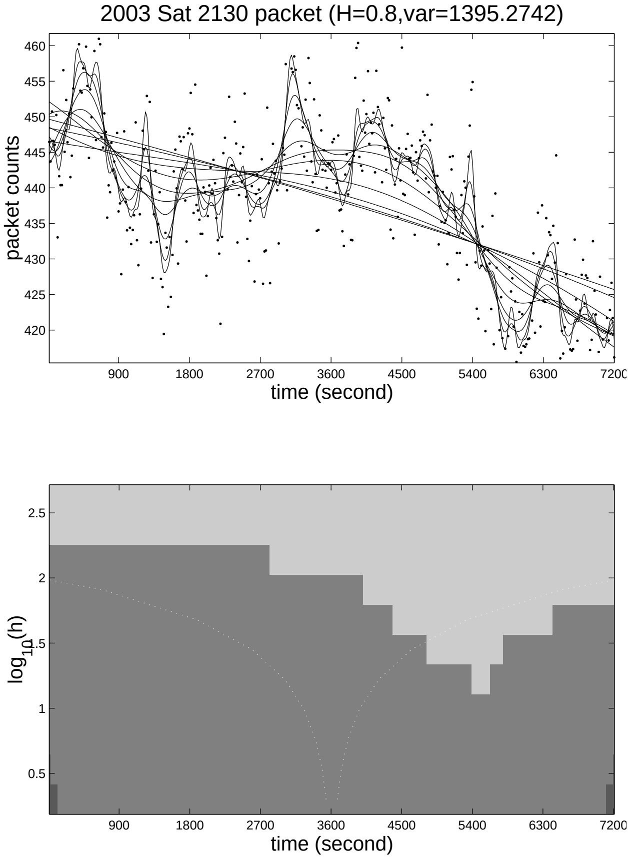


Figure 12. Dependent SiZer analysis for Saturday 9:30 PM, 2003. A statistically significant downward trend is shown.

Figure 11 shows the same analysis for data from the following year, 2003, for the same time period as Figure 9. Compared to Figure 9, a very noticeable change is the clear downward trend in the coarsest scale smooths (which are close to being lines). This trend is seen to be statistically significant by the large amounts of light shade in the top rows of the map. Perhaps this downward trend represents an important difference in either user or network behavior between 2002 and 2003. Figure 12 provides a similar analysis for 2003 for the same time slot as Figure 10. While the family of smooths appears similar between 2002 and 2003, once again there is a significant downward trend, representing a major change in user or network behavior.

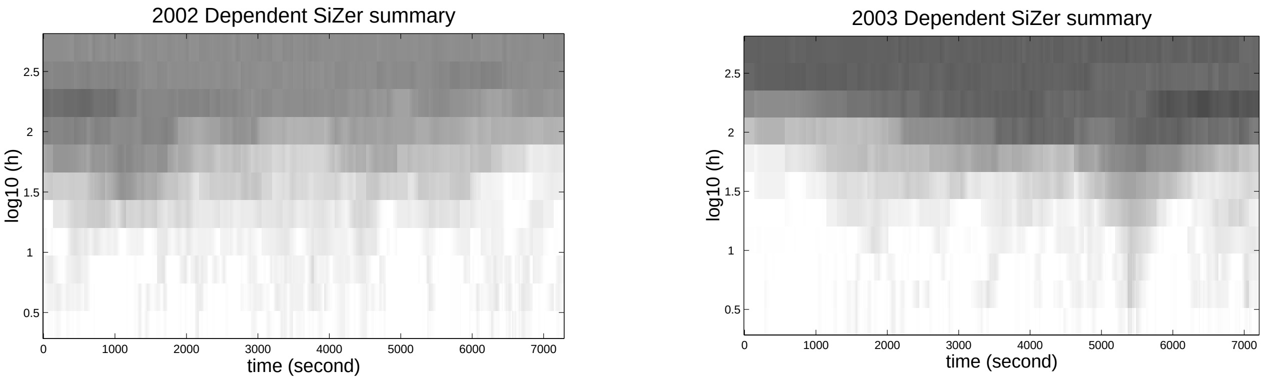


Figure 13. Dependent SiZer analysis summary for all traces in 2002 and 2003. Roughly equal amounts of non-stationary burstiness not explained by FGN existed in both years. More significant downward trends occurred in 2003.

We did a Dependent SiZer analysis for all the time series. The main observations were generally similar to what is illustrated above: there were time periods where the FGN model is appropriate (*i.e*., the map is all gray), and periods where it is not (dark and light features show up in the map). A summary of these, which allows some comparison of 2002 with 2003 is shown in Figure 13. This is a gray-level map, where the pixels represent scale and time using the same coordinates as in the individual maps. Each pixel has a gray level which represents the number of times that it was significant (either dark or light) over the entire collection. Darker shades of gray indicate more frequent significance. Comparison of the Dependent SiZer summaries for 2002 (left) and for 2003 (right) shows roughly equal amounts of non-stationary burstiness, of a type that cannot be explained by FGN, for both years. The only difference is that there were more statistically significant trends at the coarsest scales for 2003, which is quite consistent with what was observed in Figures 9-12. This suggestion that levels of burstiness (above Fractional Gaussian Noise) were similar for 2002 and 2003 is a sharp contrast with the results suggested by the Hurst parameter analysis, where there was an indication of significantly less burstiness (smaller H values) in 2003 than in 2002. Therefore, we have to look elsewhere for explanations of the shift in H between 2002 and 2003. Because we found a sharp increase in UDP packet counts between 2002 and 2003, we did a protocol-dependent analysis of the time series as described in the following section.

6. Protocol-Dependent Analysis

We processed the traces to extract individual time series for packet and byte counts for the two major transport protocols of the Internet, TCP and UDP. These protocols can influence packet and byte arrivals, especially at small time scales (*e.g*., less than one second) [12], [22]. TCP mediates the data transmission rates that applications can achieve through window-based flow- and congestion-control mechanisms [17]. UDP provides no flow or congestion control so applications using UDP must implement their own mechanisms. At larger time scales (above one second), traffic sources that have heavy-tailed sizes of application-defined data units (*e.g*., file sizes) which lead to heavy-tailed transmission "ON" times, or that have heavy-tailed durations of "OFF" times, have been shown to be the basic causes of long-range dependence in packet counts from a link that aggregates the packets of many such sources [8], [30].

We filtered each trace to extract a time series of packet or byte counts in 10 millisecond intervals for TCP packets only and likewise for UDP packets only. Because the link utilization was quite low for all traces and we were using 10 millisecond intervals, there should have been little influence of TCP packets on UDP packets and *vice versa*[3]. We then computed the logscale diagram for each of the TCP and UDP packet- and byte-count time series. Figure 14 shows the logscale diagrams for TCP and UDP along with the logscale diagrams for the complete packet time series containing both TCP and UDP ("All"). In this figure, all 28 time series for each year are overlaid in one plot. For both packet and byte counts in 2002 we found that the logscale diagram for "All" in each trace corresponded almost exactly to the one for TCP extracted from that trace. Each of the UDP logscale diagrams is completely distinct from its corresponding "All" or TCP trace. This was also the case for byte counts in 2003. Clearly, in these three cases UDP and UDP-based applications had little or no influence on the overall long-range dependence and Hurst parameter estimates reported above.

[3] At the 1Gbps speed of the monitored link, 2,500 packets of the average size 500 bytes can be transmitted in 10 milliseconds

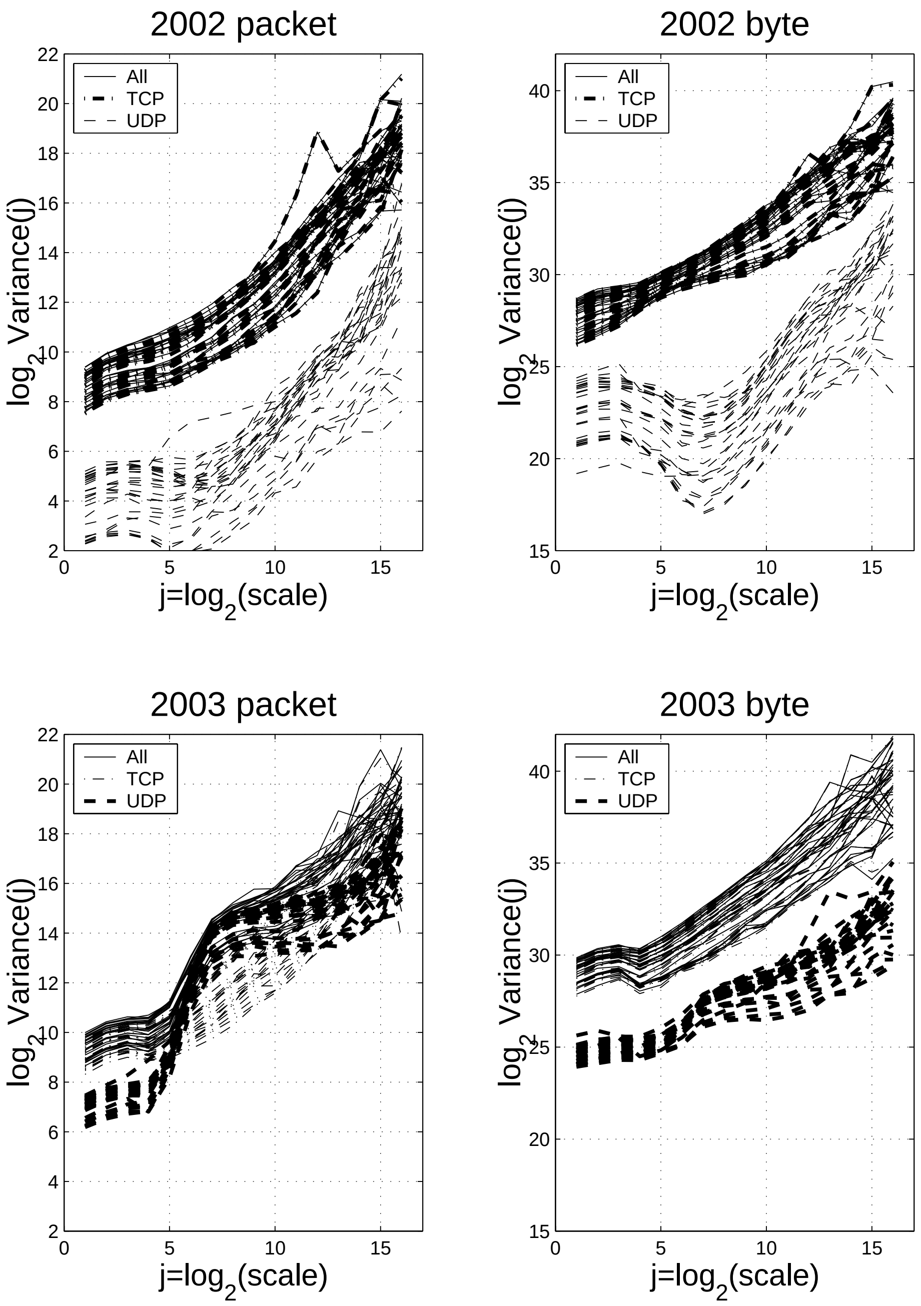


Figure 14. Logscale Diagrams for 2002 and 2003 packet and byte Counts (all traces). The plot in the lower left panel shows that UDP dominated the wavelet spectrum for all 2003 packet-count time series. The plots in the other panels show that TCP was dominant for all the other time series

There was, however, a dramatically different result for 2003 packet counts. Notice that in the lower left plot in Figure 14, the logscale diagram for “All” packet counts follows closely the shape of UDP especially in the time scales between 2^5 and 2^{10}

(300 milliseconds to 10 seconds). Figures 15 and 16 illustrate this effect in more detail for two of the representative traces (Saturday at 9:30 PM in 2002 and 2003) that were introduced earlier. .

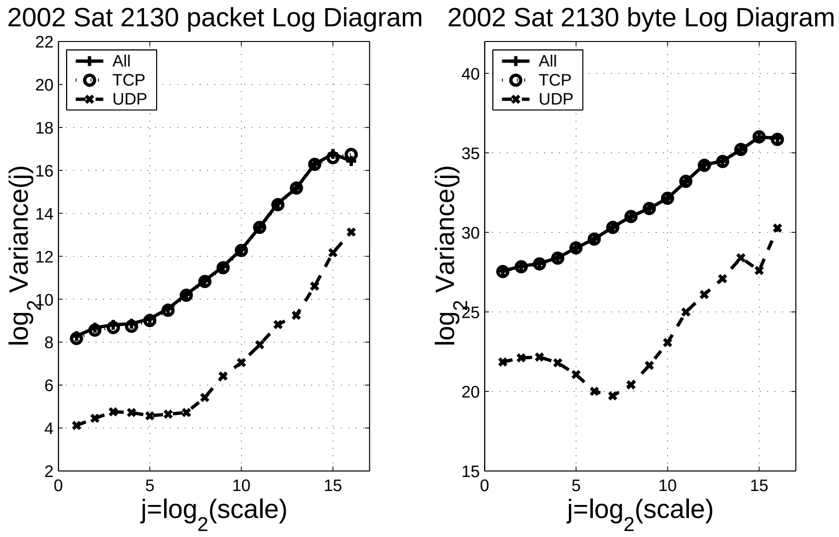


Figure 15. Logscale Diagram for Saturday at 9:30 PM, 2002. TCP dominated the wavelet spectrum for the packet and byte-count time series.

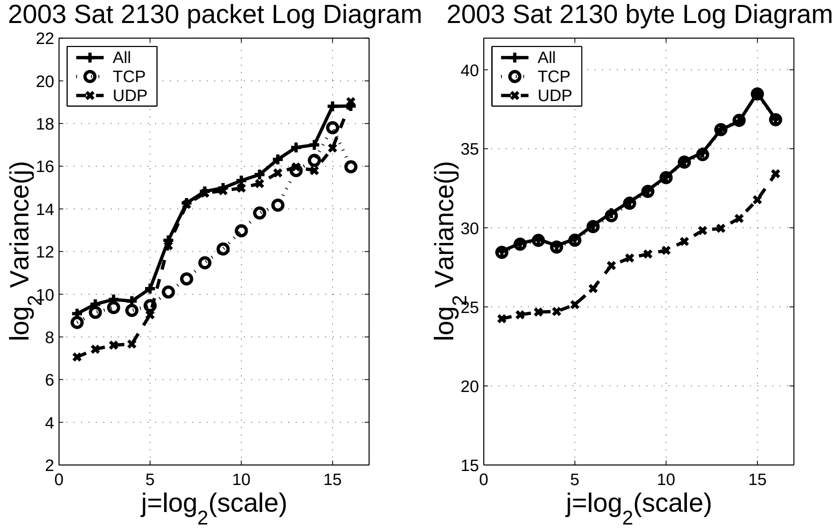


Figure 16. Logscale Diagram for Saturday at 9:30 PM, 2003. UDP dominated the wavelet spectrum for packet counts but TCP dominated for byte counts.

The steep increase in the variance of wavelet coefficients at these time scales (see Figure 16) happens near the beginning of the linear scaling region used to estimate the Hurst parameter. This implies that the slope will be smaller with a correspondingly smaller Hurst parameter estimate for packet counts. This explains the surprisingly smaller Hurst parameter estimates for 2003 as shown in Figure 1. In order to explain the reasons for these changes in UDP packet traffic between 2002 and 2003 and how they affect H, we examined the UDP traffic in more detail as described in the following.

Our 2003 traces showed a much larger percentage of UDP packets and the UDP packets exhibited two interesting characteristics not present in the 2002 traces:

- Most UDP packets had very small payloads (20-300 bytes).
- Port number 41,170 was either the source or the destination of a large fraction (*e.g*., 70%) of the UDP packets in each trace.

Port number 41,170 is known to correspond to *Blubster* (*a.k.a*. *Piolet*) [3], a music file-sharing application that became quite popular in early 2003. This suggests that the network behavior of this application was responsible for the remarkable change in the nature of UDP traffic in our 2003 traces, and this motivated us to have a deeper look at its characteristics. *Blubster* uses a proprietary protocol, so its network behavior is not documented. We examined numerous traces and a working copy of the program and found that hosts running *Blubster* (*i.e*., peers) use UDP for constructing and maintaining a network of neighboring peers and for performing searches. As an example of the use of this protocol, between 3 and 4 PM in the Thursday trace, we observed UDP packets going to more than 3,000 IP addresses in the UNC campus sent from almost 40,000 IP addresses outside it.[4] Peers also use TCP but only for downloading web pages and banners, for performing the actual file transfers, and for other housekeeping functions.

The way *Blubster* uses UDP creates an unusual pattern of burstiness in traffic. Each member of the *Blubster* network sends a significant number (40 to 125) of *ping* packets each second, with the purpose of updating its list of neighbors. The destination port number of each of these ping packets is 41,170, and they have very small payloads (20 bytes). Active neighbors reply to each ping with another small UDP packet. We also observed that *Blubster* peers continued pinging our test peer even hours after the program was closed (with the number of arrivals slowly decreasing over time). Likewise, *Blubster's* search mechanism employs UDP to exchange queries and responses using a flooding algorithm similar to one implemented in *Gnutella*. Query packets contain little more than a search string, so their payload is generally very small too. If a peer has a file matching the search string, it replies with a UDP packet that includes information about the quality of the file and the connectivity of the peer in which it is located. This reply packet is slightly larger than the other types of *Blubster* packets, but it is still relatively small (100-300 bytes). *Blubster* makes no use of UDP other than the ping and search operations. Given that this application uses very small packets, its impact is negligible in the time series of byte counts but very significant for packet counts.

Next, we analyze time series of packet and byte counts for *Blubster* traffic. We obtained these time series by first filtering packet header traces for UDP packets with one port being 41,170 and the other one being greater than 1023. The second port number condition is needed to filter out legitimate packets from other applications, such as DNS, that use well-known port numbers (0-1023, [15]). These applications can make use of UDP port 41,170 as a client port, but *Blubster* does not use well-known port numbers for its clients..

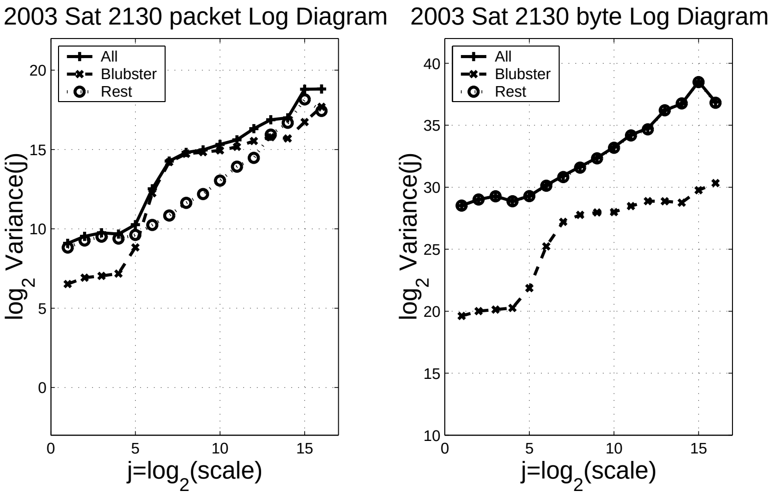


Figure 17. Logscale Diagram for Saturday at 9:30 PM, 2003. *Blubster* dominated the wavelet spectrum for packet counts but had no effect on byte counts.

[4] Probably there were fewer than 3,000 active users of *Blubster* during this interval because external peers are slow to purge their peer lists and continue to use IP addresses that may not be currently running *Blubster* or have been reassigned by DHCP.

We filtered the 2003 traces based on the *Blubster* port number to extract time series of packet and byte counts in 10 millisecond intervals for *Blubster* only and for all TCP and UDP minus *Blubster* ("Rest"). The logscale diagrams for these two time series, along with the original time series ("All"), from one representative trace (Saturday at 9:30 PM) from 2003 are shown in Figure 17. Clearly, the logscale diagram for the packet counts in the original time series ("All") is completely driven by the *Blubster* packets in the time scales between 2^5 and 2^{10}. Note that while there is a similar shape in the logscale diagram for *Blubster* byte counts, it is not reflected in the logscale diagram for the "All" byte counts. This is probably because this application uses a large number of small packets and contributes substantially less to the total bytes. We recomputed the logscale diagrams for 2003 packet counts using only the non-*Blubster* packets (the "Rest" time series) and used them to estimate the corresponding Hurst parameter. We found that these new Hurst estimates were very comparable to those from 2002, mostly falling in the range [0.90, 0.99]. For example, the Hurst estimates and confidence intervals for "Rest" in the two representative traces were 0.94 with CI of [0.92, 0.96] and 0.94 [0.90, 0.99] instead of the estimates of 0.77 [0.75, 0.79] and 0.76 [0.71, 0.81] computed from the original ("All").

We used the original SiZer method[5] to investigate why the *Blubster* application produced such distinctive scaling behavior in the logscale diagram. Figure 18 shows the SiZer analysis for one representative trace considering only the *Blubster* packets. We show three levels of detail; the full time series, a zoomed-in view of a 40 second interval at approximately 1000 seconds into the trace, and a zoomed-in view of a 10 second interval between 15 and 25 seconds from that 40 second interval. The SiZer views reveal a statistically significant (relative to white noise) high-frequency variability in the *Blubster* traffic with periods in the 1-5 second range, consistent with the time scales with heightened variance in the logscale diagram. Thus it is clear that a single peer-to-peer application with a particular UDP-based flooding-search algorithm strongly influenced the estimated Hurst parameter for the entire time series.

[5] The original version tests for structure significantly different from white noise.

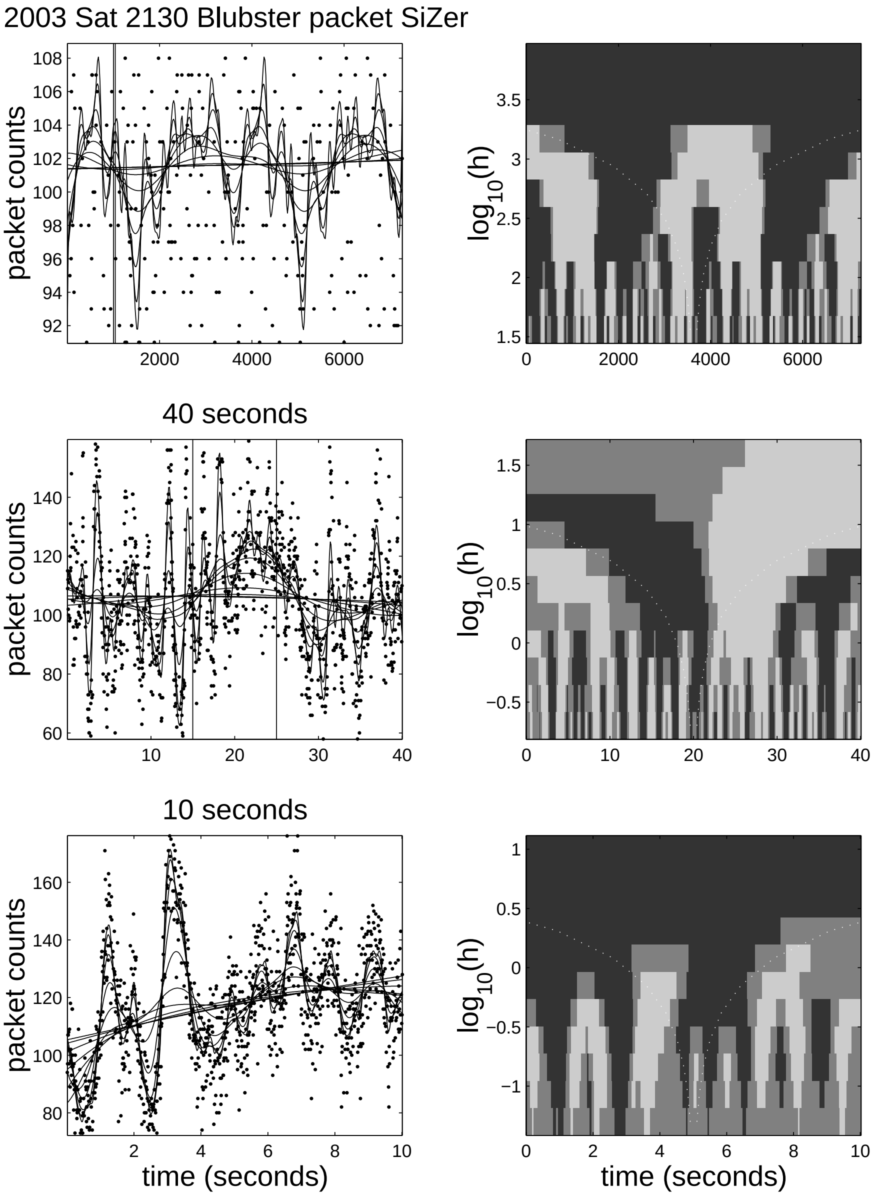


Figure 18. SiZer of Saturday 9:30 PM, 2003, Blubster packet counts. Blubster has statistically significant high-frequency variability in the 1 to 5 second time scales.

7. Summary of Results

We have presented the results from a large-scale analysis of two-hour TCP/IP packet traces acquired in 2002 and 2003 from a 1 Gbps Ethernet access link connecting the entire University of North Carolina at Chapel Hill to its Internet service provider. For those researchers interested in the long-range dependence of Internet traffic, our work has some important implications.

It should be clear from the results that depending solely on the estimated Hurst parameter as a single metric for traffic "burstiness" has several drawbacks. These include the presence of several forms of non-stationarity that cause estimates for H > 1 or very wide confidence intervals. While these phenomena exist, they may be highly localized in time (a few seconds) but obscure the H estimate for a much longer time series. Another issue is that one or two applications may strongly influence the estimate for H, especially if their traffic has a strongly periodic arrival pattern. We have demonstrated the value of employing a suite of statistical tools that includes wavelet analysis and Dependent SiZer to fully explore the nature of long-range dependence in Internet traffic and look beyond a single value for H to examine underlying structure in traffic arrival processes.

Our major findings based on the UNC traces are:

- Hurst parameter estimates for the packet counts in 10 millisecond intervals decreased significantly between 2002 and 2003 while the Hurst parameter estimates for byte counts remained about the same.
- A single peer-to-peer file sharing application (*Blubster*) that suddenly became popular between 2002 and 2003 had a strong influence on the Hurst parameter for the packet-count time series and caused the differences in H between the two years. This happened because the application sends a large number of small packets with high frequency that, in turn, produced high variance at certain time scales.
- Moderate to strong long-range dependence (Hurst parameter estimate in the range [0.65, 0.99]) existed in the vast majority of the time series.
- In several of the 56 traces studied, we found evidence of highly bursty traffic in packet or byte counts that could not be modeled as Fractional Gaussian Noise.
- The long-range dependence, as measured by the Hurst parameter estimate, was not influenced by the day-of-week or time-of-day. This implied that changes in the *active* user population (and, by implication, types of applications used) did not influence long-range dependence.
- The long-range dependence as measured by the Hurst parameter estimate was not influenced by the traffic load (link utilization) or the number of active TCP connections. In particular we found no indication of declining long-range dependence as more active TCP connections are aggregated.

.

8. References

[1] P. Abry and D. Veitch. Wavelet analysis of long-range dependent traffic. *IEEE Trans. on Information Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.

[2] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch. Self-similarity and long-range dependence through the wavelet lens. *Long-range Dependence: Theory and Applications*, P. Doukhan, G. Oppenheim, and M. S. Taqqu, eds., Birkhauser, 2000.

[3] Blubster Music Network, `http://www.blubster.com`.

[4] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. On the nonstationarity of Internet traffic. *Proc. ACM SIGMETRICS*, pp. 102–112, June 2001.

[5] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. The effect of statistical multiplexing on internet packet traffic: Theory and empirical study. *Bell Labs Technical Report*, 2001.

[6] J. Cao, W. Cleveland, D. Lin, and D. Sun. Internet traffic tends toward Poisson and independent as the load increases. *Nonlinear Estimation and Classification*, C. Holmes, D. Denison, M. Hansen, b. Yu, and B. Mallick, eds., Springer, New York, 2002.

[7] P. Chaudhuri and J. S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, vol. 94, pp. 807–823, 1999.

[8] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Trans. on Networking*, vol. 5, pp. 835–846, Dec. 1997.

[9] A. Erramilli , O. Narayan and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. on Networking*, vol. 4, no. 2, pp. 209–223, April 1996.

[10] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1996.

[11] A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades: investigating the multifractal nature of Internet WAN traffic. *Proc. ACM SIGCOMM*, pp. 25–38, Vancouver, B.C., 1998.

[12] A. Feldmann, A. C. Gilbert, P.Huang, and W. Willinger, Dynamics of IP traffic: a study of the role of variability and the impact of control. *Proc. ACM SIGCOMM*, pp. 301–313, Boston, MA, 1999.

[13] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. *ACM SIGCOMM Computer Communication Review*, vol. 26, no. 4, pp. 15–24, Oct. 1996.

[14] C. Park, F. Hernandez Campos, L. Le, J. S. Marron, J. Park, V. Pipiras, F. D. Smith, R. L. Smith, M. Trovero, and Z. Zhu. Long range dependence analysis of Internet traffic. (in submission), 2004.

[15] Internet Assigned Numbers Authority, `http://www.iana.org/assignments/port-numbers`.

[16] Internet Traces at NLANR, `http://pma.nlanr.net/`.

[17] V. Jacobson. Congestion avoidance and control. *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4, pp. 314–329, August 1988.

[18] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, vol. 91, pp. 401–407, 1996.

[19] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A Nonstationary Poisson View of Internet Traffic. *Proc. IEEE Infocom*, 2004.

[20] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. on Networking*, vol. 2, no.1, pp. 1–15, Feb. 1994.

[21] J. S. Marron. A personal view of smoothing and statistics. *Statistical Theory and Computational Aspects of Smoothing*, W. Härdle and M. Schimek, eds., pp. 1–9 (with discussion, and rejoinder 103–112), 1996.

[22] K. Park, G. Kim, and M. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. *Proc. IEEE International Conference on Network Protocols*, pp. 171–180, October 1996.

[23] K. Park, G. Kim, and M. E. Crovella. On the effect of traffic self-similarity on network performance. *Proc. SPIE International Conference on Performance and Control of Network Systems*, pp. 296-310, 1997.

[24] K. Park and W. Willinger, eds., *Self-Similar Network Traffic and Performance Evaluation*, Wiley Interscience, 2000.

[25] C. Park, J. S. Marron, and V. Rondonotti. Dependent SiZer: goodness of fit tests for time series models, *Journal of Applied Statistics*, 2004 (to appear).

[26] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[27] D. Veitch and P. Abry. A wavelet based joint estimator for the parameters of LRD. *IEEE Trans. Info. Th.*, vol. 45, no. 3, April 1999.

[28] D. Veitch, and P. Abry, Code for the estimation of Scaling Exponents, `http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html`.

[29] D. Veitch, P. Abry, and M. Taqqu. On the automatic selection of the onset of scaling. *Fractals*, vol. 11, pp. 377-390, 2003.

[30] W. Willinger , M. S. Taqqu , R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, Feb. 1997.

[31] Z.-L. Zhang, V. Ribeiro, S. Moon, and C. Diot. Small-Time scaling behaviors of Internet backbone traffic: An Empirical Study. *Proc. IEEE Infocom*, San Francisco, March, 2003.

Appendix A. Summary Data for 2002 and 2003 Traces

Tables II and III provide complete summary data for the 2002 and 2003 traces, respectively. The tables are organized with one row for each two-hour trace that gives the day and time the trace started, information about the packet and byte counts in the trace, the mean link utilization during the traced interval, the estimated number of TCP connections active per minute, and the % of packets reported as being lost by the monitor. For each of packet and byte counts, we show the total number in the trace, the % TCP, UDP and other (*e.g.*, ICMP, ARP), and the estimated Hurst parameter with confidence intervals for that time series (aggregated in 10 millisecond intervals).

UNC 2002	**Packets**					**Bytes**					**% Util.**	**Active TCP**	**% Loss**
	Total (M)	**%**			**Hurst Param. and 95% C.I.**	**Total (GB)**	**%**			**Hurst Param. and 95% C.I.**			
		TCP	**UDP**	**Rest**			**TCP**	**UDP**	**Rest**				
Sun 0500	118.8	97.87	1.89	0.24	1.04 [0.97, 1.11]	30.0	97.69	2.22	0.09	1.10 [0.90, 1.30]	3.30	11,732	0.01
Sun 1000	131.2	97.81	1.92	0.27	1.00 [0.93, 1.07]	30.5	97.02	2.86	0.12	0.76 [0.75, 0.77]	3.36	16,074	0.01
Sun 1500	185.6	95.05	4.70	0.25	0.95 [0.92, 0.98]	59.5	95.16	4.76	0.08	0.97 [0.90, 1.04]	6.53	26,890	0.06
Sun 2130	191.5	93.89	5.88	0.23	0.96 [0.93, 0.99]	60.3	94.30	5.63	0.07	0.97 [0.94, 1.00]	6.61	28,975	0.09
Mon 0500	118.5	97.82	1.86	0.32	1.12 [1.05, 1.19]	24.0	96.66	3.18	0.16	1.12 [1.00, 1.23]	2.64	11,506	0.01
Mon 1000	221.1	92.56	7.18	0.26	0.95 [0.92, 0.98]	76.9	88.03	11.86	0.11	0.98 [0.95, 1.01]	8.42	37,722	0.13
Mon 1500	243.3	93.12	6.58	0.30	0.90 [0.86, 0.95]	93.8	90.90	8.98	0.12	0.91 [0.89, 0.93]	10.26	45,126	0.14
Mon 2130	206.5	93.26	6.53	0.21	0.98 [0.93, 1.03]	67.6	94.50	5.43	0.07	0.95 [0.93, 0.97]	7.40	34,141	0.11
Tue 0500	111.9	97.77	1.94	0.29	1.06 [1.01, 1.10]	24.7	97.07	2.80	0.13	0.90 [0.83, 0.97]	2.73	11,534	0.01
Tue 1000	209.6	92.17	7.54	0.29	0.87 [0.85, 0.89]	80.8	89.04	10.87	0.09	0.87 [0.85, 0.88]	8.85	38,518	0.12
Tue 1500	253.6	91.79	7.86	0.35	0.92 [0.89, 0.95]	95.9	89.62	10.26	0.12	0.92 [0.90, 0.94]	10.48	44,818	0.16
Tue 2130	194.8	93.74	6.04	0.22	1.03 [0.98, 1.07]	60.0	94.39	5.53	0.08	0.90 [0.87, 0.93]	6.58	32,231	0.09
Wed 0500	120.9	98.11	1.63	0.26	0.99 [0.94, 1.04]	23.9	97.33	2.55	0.12	1.09 [0.97, 1.20]	2.63	11,172	0.00
Wed 1000	224.7	91.94	7.80	0.26	0.84 [0.44, 1.24]	86.9	87.50	12.41	0.09	0.65 [0.26, 1.05]	9.50	38,173	0.14
Wed 1500	245.0	91.34	8.38	0.28	1.07 [1.00, 1.14]	91.8	87.48	12.41	0.11	1.23 [1.11, 1.34]	10.04	43,098	N/A
Wed 2130	212.5	93.69	6.08	0.23	0.98 [0.94, 1.03]	76.1	93.21	6.72	0.07	0.96 [0.93, 0.99]	8.34	32,390	0.12
Thu 0500	113.2	97.18	2.50	0.32	1.10 [1.03, 1.17]	27.3	97.11	2.77	0.12	0.91 [0.84, 0.98]	3.01	11,002	0.01
Thu 1000	209.6	90.85	8.83	0.32	0.90 [0.88, 0.92]	86.3	87.44	12.45	0.11	1.00 [0.95, 1.04]	9.46	38,555	0.13
Thu 1500	221.2	91.13	8.57	0.30	0.92 [0.87, 0.96]	83.6	88.39	11.51	0.10	0.86 [0.85, 0.87]	9.15	43,020	0.11
Thu 2130	178.0	91.49	8.20	0.31	1.00 [0.95, 1.04]	50.8	93.03	6.86	0.11	0.96 [0.91, 1.00]	5.58	29,763	0.06
Fri 0500	110.2	94.34	5.29	0.37	0.96 [0.91, 1.00]	23.5	94.06	5.79	0.15	0.86 [0.79, 0.93]	2.59	11,843	0.00
Fri 1000	208.9	91.49	8.24	0.27	0.95 [0.92, 0.98]	76.6	86.70	13.19	0.11	0.91 [0.89, 0.93]	8.40	38,026	0.12
Fri 1500	214.7	90.95	8.63	0.42	1.06 [1.02, 1.11]	79.8	88.58	11.17	0.25	0.94 [0.92, 0.96]	8.74	39,856	0.10
Fri 2130	161.3	92.89	6.77	0.34	0.97 [0.92, 1.01]	44.5	94.34	5.54	0.12	0.86 [0.84, 0.88]	4.89	21,482	0.04
Sat 0500	108.1	94.37	5.19	0.44	1.06 [0.94, 1.18]	25.0	96.39	3.45	0.16	0.88 [0.81, 0.95]	2.75	8,246	0.00
Sat 1000	139.9	94.65	5.02	0.33	0.91 [0.88, 0.94]	38.4	95.79	3.97	0.24	0.79 [0.78, 0.79]	4.22	21,409	0.02
Sat 1500	168.8	93.14	6.50	0.36	0.97 [0.92, 1.01]	55.7	92.61	7.25	0.14	0.93 [0.88, 0.97]	6.12	22,032	0.04
Sat 2130	149.9	92.81	6.89	0.30	0.98 [0.93, 1.02]	46.9	94.63	5.28	0.09	0.84 [0.83, 0.85]	5.15	18,483	0.03

Table II. Trace Summary Data for 2002

UNC 2003	Packets					Bytes					% Util.	Active TCP	% Loss
	Total (M)	%			Hurst Param. and 95% C.I.	Total (GB)	%			Hurst Param. and 95% C.I.			
		TCP	UDP	Rest			TCP	UDP	Rest				
Sun 0500	233	74.73	25.00	0.27	0.72 [0.65, 0.79]	48.2	84.09	15.80	0.11	0.90 [0.88, 0.92]	5.35	12,840	0.00
Sun 1000	268	69.11	30.56	0.33	0.90 [0.78, 1.01]	50.8	79.55	20.28	0.17	0.90 [0.89, 0.92]	5.65	20,241	0.00
Sun 1500	342	67.79	31.92	0.29	0.71 [0.69, 0.73]	95.7	84.81	15.09	0.10	0.95 [0.92, 0.98]	10.63	32,466	0.00
Sun 2130	361	71.29	28.42	0.29	0.71 [0.68, 0.73]	117.8	88.61	11.30	0.09	0.90 [0.89, 0.92]	13.09	37,793	0.00
Mon 0500	233	77.93	21.78	0.29	0.93 [0.81, 1.05]	51.0	86.34	13.51	0.15	0.87 [0.84, 0.90]	5.67	14,795	0.00
Mon 1000	394	76.04	23.69	0.27	1.31 [1.13, 1.49]	128.8	85.02	14.87	0.11	0.97 [0.96, 0.98]	14.31	48,252	0.00
Mon 1500	449	73.48	26.34	0.18	0.81 [0.76, 0.85]	159.5	87.01	12.94	0.05	0.93 [0.92, 0.95]	17.72	56,839	0.00
Mon 2130	368	74.00	25.73	0.27	0.80 [0.75, 0.84]	124.0	89.21	10.72	0.07	1.00 [0.96, 1.05]	13.78	39,127	0.00
Tue 0500	254	80.65	18.94	0.41	0.76 [0.69, 0.83]	52.1	86.88	12.96	0.16	0.80 [0.73, 0.87]	5.79	16,871	0.00
Tue 1000	406	77.55	22.07	0.38	0.69 [0.66, 0.72]	133.8	85.73	14.17	0.10	0.90 [0.89, 0.92]	14.86	49,494	0.00
Tue 1500	455	74.91	24.77	0.32	0.72 [0.69, 0.75]	153.2	85.49	14.41	0.10	0.92 [0.90, 0.94]	17.02	56,913	0.00
Tue 2130	395	72.72	26.82	0.46	0.78 [0.73, 0.82]	123.1	88.53	11.34	0.13	0.90 [0.88, 0.91]	13.68	43,276	0.00
Wed 0500	250	78.94	20.44	0.62	0.65 [0.62, 0.68]	53.6	87.18	12.60	0.22	0.85 [0.81, 0.90]	5.96	16,671	0.00
Wed 1000	401	77.59	22.05	0.36	0.71 [0.69, 0.73]	127.9	84.64	15.26	0.10	0.93 [0.92, 0.94]	14.21	49,732	0.00
Wed 1500	456	74.84	24.78	0.38	0.88 [0.80, 0.95]	164.3	86.49	13.40	0.11	0.92 [0.91, 0.93]	18.26	59,064	0.00
Wed 2130	376	71.33	28.19	0.48	0.72 [0.69, 0.75]	118.2	88.16	11.64	0.20	0.89 [0.88, 0.90]	13.13	42,549	0.00
Thu 0500	250	79.67	19.82	0.51	0.92 [0.85, 0.99]	54.6	87.87	11.87	0.26	0.90 [0.87, 0.93]	6.06	15,697	0.00
Thu 1000	389	76.77	22.88	0.35	0.79 [0.74, 0.83]	130.4	84.93	14.96	0.11	0.93 [0.92, 0.95]	14.49	45,785	0.00
Thu 1500	456	74.39	25.25	0.36	0.77 [0.75, 0.79]	159.6	86.21	13.64	0.15	0.96 [0.95, 0.98]	17.73	55,104	0.00
Thu 2130	379	70.65	28.99	0.36	0.75 [0.73, 0.77]	118.8	87.00	12.86	0.14	0.94 [0.93, 0.95]	13.20	37,398	0.00
Fri 0500	257	76.80	22.71	0.49	0.82 [0.74, 0.89]	52.8	85.02	14.71	0.27	0.94 [0.92, 0.96]	5.87	16,674	0.00
Fri 1000	412	75.33	24.33	0.34	0.73 [0.71, 0.75]	137.2	84.88	14.98	0.14	0.94 [0.93, 0.95]	15.25	48,933	0.00
Fri 1500	443	73.16	26.54	0.30	0.79 [0.78, 0.79]	153.3	86.08	13.79	0.13	0.87 [0.46, 1.27]	17.04	48,869	0.00
Fri 2130	335	72.09	27.56	0.35	0.71 [0.69, 0.73]	98.1	87.73	12.13	0.14	1.01 [0.96, 1.05]	10.90	25,540	0.00
Sat 0500	241	74.47	25.06	0.47	0.78 [0.71, 0.85]	50.9	84.09	15.73	0.18	0.93 [0.88, 0.98]	5.66	12,824	0.00
Sat 1000	290	69.69	29.92	0.39	1.18 [0.98, 1.39]	62.1	81.86	17.93	0.21	0.87 [0.86, 0.88]	6.90	23,087	0.00
Sat 1500	340	69.61	30.02	0.37	0.83 [0.76, 0.91]	95.1	85.66	14.23	0.11	0.87 [0.84, 0.90]	10.56	27,332	0.00
Sat 2130	316	69.34	30.30	0.36	0.76 [0.71, 0.81]	78.4	84.87	15.01	0.12	0.89 [0.88, 0.90]	8.71	24,003	0.00

Table III. Trace Summary Data for 2003