Capturing the Elusive Poissonity in Web Traffic

C. Park
Dept. of Statistics
Univ. of Georgia
Athens, GA 30602-1952, USA
cpark@stat.uga.edu

H. Shen J. S. Marron
Dept. of Statistics and Op. Research
Univ. of North Carolina
Chapel Hill, NC 27599-3260, USA
{*haipeng,marron*}*@email.unc.edu*

F. Hernández-Campos*
Dept. of Computer Science
Univ. of North Carolina
Chapel Hill, NC 27599-3175, USA
fhernand@cs.unc.edu

D. Veitch
CUBIN†
Dept. of Electrical and Electronic Eng.
Univ. of Melbourne, Australia
d.veitch@ee.unimelb.edu.au

Abstract

Numerous studies have shown that the process of packet arrivals from Web traffic exhibits strong long-range dependence, which makes it not amenable to be described using the convenient but necessarily short-range dependent framework of Poisson modeling. However, Web traffic is ultimately driven by independent human behavior, so it seems natural to search for an underlying "seed process", consistent with Poissonity, indirectly driving the packet arrivals of Web traffic. Our study examines Web traffic at different levels of packet aggregation, using powerful statistical analysis tools for identifying the finest level that can be effectively modeled using a homogeneous Poisson process. We show that the arrivals of HTTP responses, TCP connections and Web pages do not provide a satisfactory seed process. However, we find Poissonity in the arrivals of "navigation bursts". A navigation burst is a tightly-spaced sequence of Web pages downloaded by the same Web client, which can be explained by fast navigation through several pages before reaching relevant content. Our analysis suggests that the start times of such navigation bursts, which we identify by detecting user think times between 12 and 30 seconds, can be effectively modeled as a homogeneous Poisson process. We believe that our methodology can be extended to other complex modeling problems where finding Poissonity can greatly simplify parsimonious modeling.

1. Introduction

Internet traffic has been the focus of numerous studies in recent years. One of most influential results in this area was the finding of pervasive long-range dependence (LRD) in the process of IP (Internet Protocol) packet arrivals on Internet links by Leland *et al.* [10, 22]. This characteristic of Internet traffic is in sharp contrast with earlier "teletraffic" modeling work in the context of telephone networks [21], where short-range dependent processes, including the memoryless Poisson process, were widely applicable. Internet traffic shows high variability across a wide range of time scales, while a Poisson process necessarily exhibits a rapid decrease in its variability as the scale of temporal aggregation is increased. This "failure of Poisson modeling", in the words of Paxson and Floyd [15], has lead to a rich literature on the modeling of Internet traffic using the mathematics of fractal and multi-fractal stochastic processes.

Interestingly, and despite more than a decade of intense work, no single traffic model (or even set of models) has emerged as the agreed reference in the field. Proposed models are often too narrowly applicable or too complicated to be used by networking practitioners, despite the ongoing need for a more formal understanding of traffic for tasks such as capacity planning and anomaly detection. Perhaps more seriously, there is typically a disconnection between the component elements of these models and the network mechanisms that lie behind them. Without physical meaning as a guide, and a related "physics" linking mechanisms with observed arrival processes, it is problematic to demonstrate why a given model is better than any other.

While the evidence against modeling the process of In-

*Currently at Google Inc., Mountain View, California.

†ARC Special Research Centre on Ultra-Broadband Information Networks (CUBIN). CUBIN is an affiliated program of National ICT Australia (NICTA).

ternet packet arrivals directly using a Poisson process is clear and overwhelming, we argue that it is nonetheless of great interest to search for *Poissonity* in the higher-level mechanisms and behaviors generating packet arrivals. In the end, Internet traffic arises from the superposition of the communications between a large numbers of hosts, which are generally independent of each other, particularly because independent human behavior is the ultimate driver of (most) applications. Although dependencies between user, application, host and network are complex, the paradigm of "many independent events, each with a small probability of occurring" should hold sway at a high enough level. This is the classical mechanism generating Poisson events. We believe that identifying which events form this "seed" Poisson behavior can lead to a physically meaningful skeleton upon which a traffic model which is both intuitive and tractable can be based. For completeness, recall that a Poisson process can be defined as the point process where inter-arrival times are both mutually independent, and exponentially distributed with the same parameter.

Our work aligns with recent efforts by Hohn *et al.* [9] to explain packet arrivals using a compound point process whose seed process is Poisson. We also consider the question of finding Poissonity in Internet traffic, but focus on Web traffic. Web traffic provides a particularly good case study. First, packet arrivals from Web traffic are known to exhibit long-range dependence, as Crovella and Bestavros demonstrated [5]. Second, even after the emergence of file-sharing applications, Web traffic represents a large fraction of the traffic on the Internet. Finally, Web traffic is a well-understood traffic type, so we can explain our findings in terms of the characteristics of Web browsing and the HyperText Transfer Protocol (HTTP) [3]. We can also re-use accepted measurement techniques to construct large and rich datasets for analysis. Although, as in [9], our ultimate goal is a physically meaningful model of the packet arrival process itself, here the lowest level object we study directly is HTTP responses, which has structure, notably long-range-dependence, which was ignored in [9] and merits further study. For completeness, recall that long-range dependence is defined as a slow, power-law divergence in the Fourier spectrum $\Gamma(\nu)$ at low frequencies (or equivalently large lag in the time domain): $\Gamma(\nu) \sim c|\nu|^{-\alpha}$, as $\nu \to 0$, where $\alpha \in (0,1)$ is the LRD exponent, and $c > 0$. In contrast, short-range dependent processes have a finite spectrum at the origin: $\Gamma(0) = \Gamma_0, 0 < \Gamma_0 < \infty$.

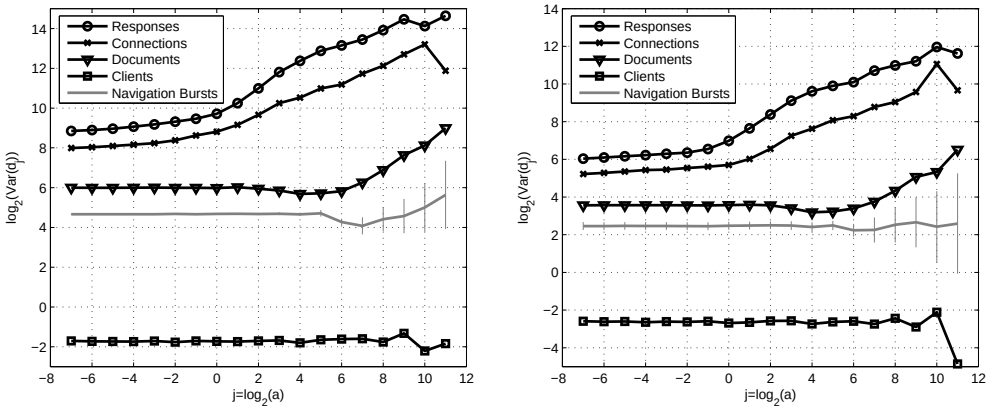
Our study examines different levels at which packets from Web traffic can be combined or aggregated, such as individual HTTP responses or entire Web pages, in the search for the finest level at which a homogeneous Poisson process emerges as a viable model. Such a model will then work as a natural "seed process" for the parsimonious modeling of packet arrival times. Our analysis relies on two statistical tools, plots of wavelet spectra [1] and SiZer maps [4], which are far more robust to non-stationarities that the variance-time plot frequently employed in the past (see [14]).

We now briefly describe wavelet spectra (see [1] and [18] for more details, the Matlab code we use is available at [19]). The (log) wavelet spectrum plots the ($\log_2$) estimated variance of the wavelet coefficients of the analyzed process as a function of the ($\log_2$) time scale j. We use Daubechies wavelets [6] with three vanishing moments. The important points to note are the following. A Poisson process has a flat spectrum (just as in the Fourier spectrum case), whereas long-range dependence manifests as a straight line behavior at large scales. All the processes studied here are arrival times, and are therefore point processes. The estimated spectrum is strongly influenced by the amount of data. Roughly speaking, doubling the length of a data set doubles the total amount of variance or energy present, resulting in a shift upward of one in the (log) spectrum. Finally, as the amount of data effectively halves with each unit increase in j, confidence intervals grow rapidly with scale.

Figure 1 provides a concise overview of our findings, using plots of the wavelet spectra from two datasets derived from packet header traces, collected over two four hour periods at the University of North Carolina at Chapel Hill (UNC) in 2001. By studying sequence and acknowledgment numbers in observed Transport Control Protocol (TCP) packets, we were able to reconstruct the arrival process of Web traffic at multiple levels of aggregation: individual responses (*e.g.*, an HTML file, an image), TCP connections, entire Web documents (*i.e.*, Web pages), navigation bursts (*i.e.*, "dense" sets of documents), and distinct clients. It is essential to note that by "aggregation", we do not mean simple block time averages, but rather grouping of packets into larger sets or objects. By analyzing the data at a given "aggregation level", we mean the analysis of the *arrival process* only of the objects at that level.

The spectra corresponding to 5 different objects in Figure 1 show roughly similar features: a flatness at small scales indicative of a Poisson-like lack of structure, deviation from the flatness beginning at some scale reflecting energy due to clustering, leading at large scale to a straight line behavior indicative of LRD (note that, to within confidence intervals, the deviations from a straight line at the very largest scales are not significant). However, depending on the aggregation level, the strength of deviations from Poisson behavior, and the range of "non-Poisson" scales, varies greatly.

At one extreme, the arrival process of clients is indistinguishable from Poisson over all scales. Intuitively this makes sense, as the client definition captures a great deal of structure, including the dependencies induced by multiple browsing sessions, TCP dynamics, as well as source characteristics, notably the heavy tailed nature of files which is



(a) 2001 Thursday (b) 2001 Sunday

Figure 1. Wavelet spectra of response, connection, document, client, and navigation burst arrivals for (a) 2001THU (b) 2001SUN: LRD clearly dominates at the response and connection levels, and still appears at the document level for coarse time-scales. Poissonity dominates at the client level and it is also valid at the navigation burst level (defined using a 12-second threshold).

the commonly accepted underlying cause of LRD in packet data. Thus, the client object is so large that the LRD is carried within it, leaving the arrival process free of it. On the other hand there is no reason in a stationary regime, and certainly no network protocol reason, why the arrival of clients should have any clustering, and so the arrivals are Poisson. However, our aim is not to find just "any" Poisson process in traffic, but to find the finest level object which is Poisson: clients are too coarse to be the parsimonious choice for modeling. Returning to Figure 1, the spectra for responses and connections are clearly non-Poisson: although responses and connections each capture a certain amount of packet structure, it is not sufficient, these objects are too "small".

Web *documents* are objects whose size is intermediate between that of responses and clients. Documents capture dependencies including how pages are constructed and retrieved, and are an obvious candidate for a "session level" aggregation [17]. From Figure 1, we see that, although the deviation from Poisson is much reduced both in amplitude and in terms of the time scales affected, they are still significant relative to confidence intervals. It seems that there are dependencies between different documents, so that they cannot play the independent seed role we seek.

One way in which Web documents could be connected is through user behavior. Users tend to click on a sequence of links before detailed reading of content. Such *navigation bursts*, essentially a tightly clustered sequence of Web document downloads from a single client, can be naturally separated from each other by *think times*. As shown in Figure 1, navigation bursts capture sufficient structure to allow their arrival process to be Poisson, yet are much smaller than clients. This aggregation level makes physical sense, since independent exponential waiting times, which characterize a homogeneous Poisson process, are a good model for human reading time, but not for structures inside the navigation bursts.

In this paper we propose navigation bursts as the natural seed process on which to base packet and response level modeling for HTTP data. As navigation bursts cannot be directly observed but only inferred from our data, timeouts are used as part of the navigation burst definition. We use a simple timeout, with value around 12 seconds. This value is not arbitrary, but is based on observations demonstrating that it is in some sense the natural scale at which Poissonity emerges. We emphasize that the significance of our work is not that we simply offer, yet another, definition of a "session" level for traffic modeling or generation. Indeed this could have been done, and even justified intuitively, without even looking at data! Instead, our contribution is that we have identified a good candidate process which, for solid empirical reasons, localizes the interface where clustering begins in the space of protocol/human interactions. Our contribution is also the unambiguous *rejection* of several alternative processes for this role, again based on a large amount of empirical evidence. We leave to future work the actual construction of a full model based on the navigation burst skeleton.

The current paper is organized as follows. Section 2 describes the data used in the current study, and defines in detail four different aggregation levels of Web traffic. Then their properties are summarized using the wavelet spectrum

Level	Specification	Data Acquisition
Response	HTTP entity, RFC 1945 [3]	Uninterrupted increase of data sequence numbers from server
Connection	TCP connection, RFC 793 [16]	TCP packets with the same IP addresses and port numbers
Document	Responses forming a Web page	Set of responses separated by 1 second of idle time
Navigation Burst	Dense sequence of documents	Set of documents separated by 12 to 30 seconds of idle time
Client	Documents browsed by one user	Set of documents downloaded from the same client IP

Table 1. Summary of the levels of aggregation at which we examine Web traffic in this paper.

and SiZer analysis. Section 3 discusses navigation bursts. We study various threshold values for the think time and determine which values capture the Poisson properties, and explain why. We conclude in Section 4.

2. Analysis of Responses, Connections, Documents and Clients

2.1. Data Description

The starting point of our study is a collection of TCP/IP header traces collected in 2001 and 2002 from the Gigabit link connecting the University of North Carolina at Chapel Hill (UNC) to the Internet. Our analysis focuses on the activity of UNC clients, which we extract from the traces by filtering for TCP connections started from the university and with port 80 as their destination. These connections represent the browsing activity of UNC's population, which includes roughly 26,000 students, (all of whom have networked computers), 3,000 faculty members and 9,000 staff members. For illustration purposes, this paper primarily examines a trace collected between 1 PM and 5 PM on Thursday, April 26, 2001 (2001THU). This trace has a high level of Web traffic activity, and it appears particularly stationary. In addition, we also report selected results from a trace collected between 8 AM and noon on Sunday, April 29, 2001 (2001SUN). We also studied another two similar datasets derived from packet header traces collected in April 2002. Our full set of results is available online [13].

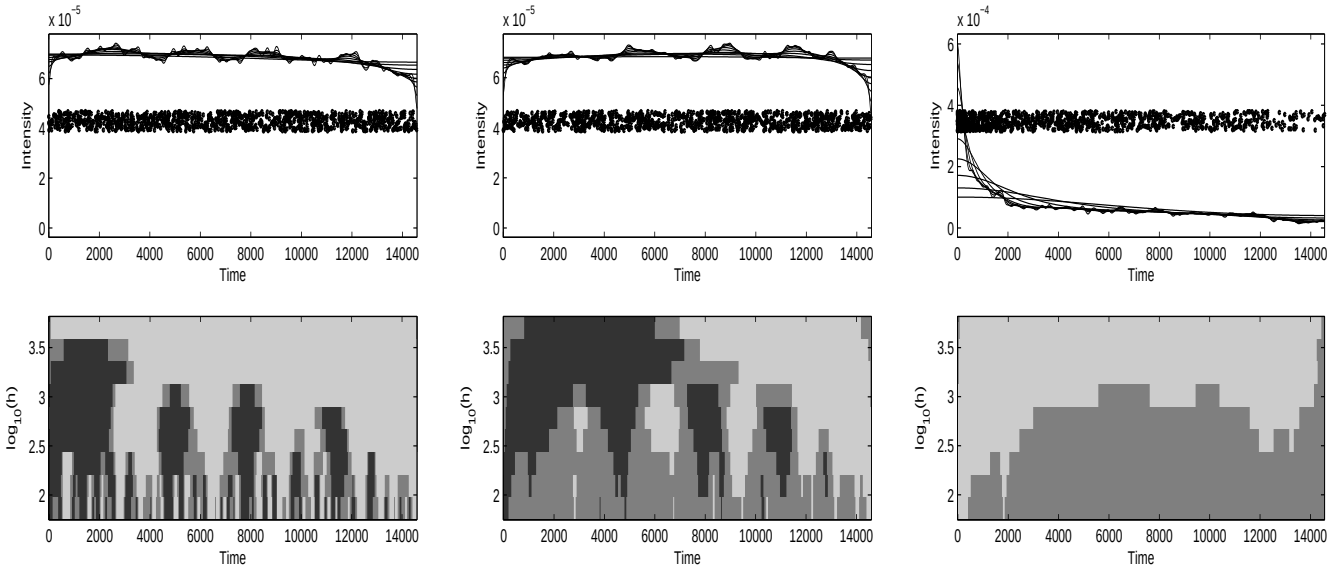
Our work considers the statistical properties of Web traffic at several levels. At the lowest level, we consider Web responses, which are the individual HTTP entities (roughly "files") downloaded from Web servers by Web browsers [3]. Web responses have been naturally grouped in two ways, by TCP connection and by Web page [17]. Finally, we also consider a client level, which we define as the entire set of responses, connections and documents with the same client IP address. These four levels of responses, connections, documents and clients create a rich data set and a challenging modeling problem. Our multi-level datasets are available online [8]. In total, we studied 19 millions HTTP responses carried in 10.5 million TCP connection, originating from 49,049 distinct clients[1]. Our multi-level analysis is more focused on the impact of human browsing on the properties of Web traffic than earlier studies. For example, Nuzman *et al.* [12] concentrated only on TCP connection arrivals in Web traffic.

For illustration purposes, Figure 2 shows the SiZer plots of the intensity estimates of the arrivals of responses, documents, and clients for the 2001THU trace, respectively. Since homogeneous Poisson processes have constant intensity functions, the idea is to check whether the estimated intensity is close to a constant function for each aggregation level.

SiZer is based on kernel density estimation (for example, see [20]) of the data, some of which are displayed as dots in the top panel of Figure 2. These estimated intensities are shown as thin curves corresponding to different window widths. This is called a family of smooths which is indexed by the window widths, and each of the thin curves represents a different row of the SiZer map in the low panel, *i.e.*, level of resolution of the data. Essentially, the top panel of Figure 2 shows the kernel intensity estimates of the arrivals of the corresponding levels. The x axis represents time in seconds (over four hours) and the y axis is the intensity. For a Poisson process, its family of smooths should be close to a constant function. Note that the present paper examines traffic at time-scales of a few hours, were diurnal effects can generally be ignored. Otherwise, the analysis must consider non-homogeneity in the Poisson process.

The lower panel of Figure 2 is called the SiZer map, which is doing a graphical statistical inference. In particular, it uses different shades of gray to flag trends in the intensity estimates that are statistically significant compared to natural variation. The horizontal axis represents time, and thus is the same as the horizontal axis in the top panel. The vertical locations correspond to the same logarithmically spaced window widths that are used for the family of intensity estimates (thin curves) in the top panel. At each scale-time location (*i.e.*, pixel in the map) statistical inference is done on the slope of the corresponding curve and the test results are reported using a color scheme. Regions shaded in dark gray in the SiZer map indicate sta-

[1]Note that anonymization prevented us from recognizing the same client across more than one trace.



(a) Response Arrivals (b) Document Arrivals (c) Client Arrivals

Figure 2. SiZer plots (2001THU) of (a) response arrivals, (b) document arrivals, and (c) client arrivals.

tistically significant increases (of the thin curves in the top panel), while light gray regions indicate statistically significant decreases. Region showing an intermediate shade of gray indicate lack of statistical significance, suggesting that the data are consistent with natural variation. For a Poisson process, its SiZer map is expected to show only intermediate gray regions.

2.2. Responses

Web responses are composed of an HTTP header and a payload. The payload can be a file, such as HTML source file or an image file, or some dynamically generated content. Since our data is extracted purely from packet header traces, we cannot learn anything about the semantics of Web response. Therefore we focus on Web responses in the abstract, which can be accurately extracted from packet header traces using the techniques described in [11, 17]. Briefly, our analysis method consists of analyzing unidirectional traces of TCP/IP headers sent from Web servers to clients (browsers) in order to infer application-level characteristics of Web traffic. In particular, we exploit properties of TCP's sequence numbers and increases in acknowledgment number to determine request and response sizes. The basic idea is to observe that Web responses are composed of one or more TCP segments with consecutive sequence numbers. If a connection has a single response, every data segment sent from the Web server to the Web browser is part of the same response. The size of the response is then given by the difference between the highest and the lowest sequence number of these segments. Similarly, the duration of the response is given by the difference between the first and the last timestamps of the segments. Every packet trace we examined in this paper included a timestamp, accurate within a few hundred microseconds, of the time at which the packet reached the monitoring point. We can therefore measure response duration, and times between responses, accurately.

Sequence number analysis is a well-known, straightforward measurement technique, and it does *not* require timeouts or heuristics of any kind. However, its implementation requires careful handling of packet reordering, retransmission and sequence number wraparound. None of these difficulties should have introduced any inaccuracies in our data. Note also that this basic technique can be extended to study TCP connections with more than one HTTP response (*i.e.*, persistent connections). The key observation is that a request must necessarily be present between two consecutive responses, so an increase in client-to-server sequence number defines a boundary between two responses.

As long as no HTTP pipelining [7] is used, this method can always identify the set of packets that form each Web response. HTTP pipelining was very uncommon in 2001 and 2002 [17], so we are highly confident on the accuracy of our method for extracting Web response data from our traces.

As mentioned above, the SiZer plot of the response arrivals for the 2001THU trace is shown in Figure 2(a). The

family of smooths in the top panel shows that the estimated intensity has many big and small oscillations, which indicates that the process is far from Poisson processes. The SiZer map in the lower panel of Figure 2(a) shows that the oscillations apparent in the top panel are statistically significant because there appear to be many dark gray (increasing) and light gray (decreasing) regions at both high (near the bottom of the SiZer map) and low (near the top) levels of resolution. The major lesson from Figure 2(a) is that the arrival process of responses does not have homogeneous Poisson properties as concluded from the wavelet spectra in Figure 1.

The exponential quantile plot of the response interarrival times suggests that they are also far from being exponential, which is not shown here to save space. We intend to construct a Poisson cluster model for the Web response arrivals in a future manuscript.

2.3. Connections

TCP connections carrying Web traffic contain one or more HTTP request/response pairs. In the first specification of HTTP (1.0) [3], a new TCP connection was required for each pair. After the connection was established, a single request was sent from the client to the server, followed by a single response sent from the server to the client, and the connection was closed. Later versions of the protocol [7] introduced the concept of persistent connections, which could remain open after the first response was transferred, and carry new request/response pairs. This eliminates the extra delay that each connection establishment involves, and also avoids the slower sending rates at the beginning of TCP connections. Our method for processing TCP/IP header traces to extract HTTP responses can easily be extended to study other TCP connection characteristics, such as their start times and the number of responses they carried. Note that our concept of connection is not the same as the connection used in [12], whose definition is more like our concept of responses. Also, our analysis not only identified individual connections using IP addresses and port number, but also detected the reuse of port numbers by examining SYN packets and their sequence numbers.

In search for the Poissonity, we also examined the arrival process of connections. We do not report its SiZer plot, and the result [13] suggests that it is very similar to that of the response arrivals, which does not imply a Poisson process as concluded from the wavelet spectra in Figure 1.

2.4. Documents

Web content is generally organized using Web documents (*i.e.,* Web pages), which consist of a base HTML source and embedded objects. Embedded objects include images, audio files, style files, *etc.* When a Web browser downloads a Web document from a Web server, it uses the first request/response pair to download the base HTML source. After receiving and parsing this source, the browser uses one request/response pair for downloading each of the embedded objects in the document. As a consequence, Web documents create significant dependencies in the arrival process of Web responses. The arrivals of documents provide a more aggregated arrival process where Poissonity could arise.

Extracting information about documents from TCP/IP header traces is more difficult than extracting information about responses. In this paper, we rely on the well-known *think time* heuristic to group responses into documents [2, 11, 17]. The starting point of this heuristic is the observation that users navigate the Web by downloading a sequence of documents. The user has to spend some time reading the content of each Web document before clicking on a link or typing a new URL in the browser. Therefore, no network activity occurs during these user think times, and this fact can be used to group responses into documents. Unfortunately, inactivity can also be due to other causes, such as network losses, processing times, *etc.*. We distinguish user think times from the other kinds of inactivity periods using a fixed threshold of 1 second. Two responses separated by more than 1 second of network inactivity are not considered to be part of the same document.

While there is some degree of uncertainty in the think time heuristic, it should provide a reasonably accurate dataset for characterizing the arrival process of Web documents. We will also use the term *think time* to refer to those inactivity periods in which no network traffic is observed for a given Web client. The duration of a think time is given by the difference between the timestamp of the last segment of the last response before the think time and the timestamp of the first segment of the first response after the think time.

This definition of documents was carefully examined in [11] and [5], who demonstrated that a think time threshold around 1 second are reasonable and provide consistent results. Unfortunately, we know of no validation of these results using actual HTTP payloads.

It is important to clarify the relationship between Web documents and TCP connections. When only non-persistent connections are used, downloading an entire Web document requires as many TCP connections as embedded objects in the document. In this case, documents provide a level of aggregation higher than that of connections. However, the use of persistent connections complicates this picture. Modern browsers use two to four persistent connections (to each Web server) to download a Web document, and each connection carries one or more request/response pairs. If a second document is downloaded from the same Web server,

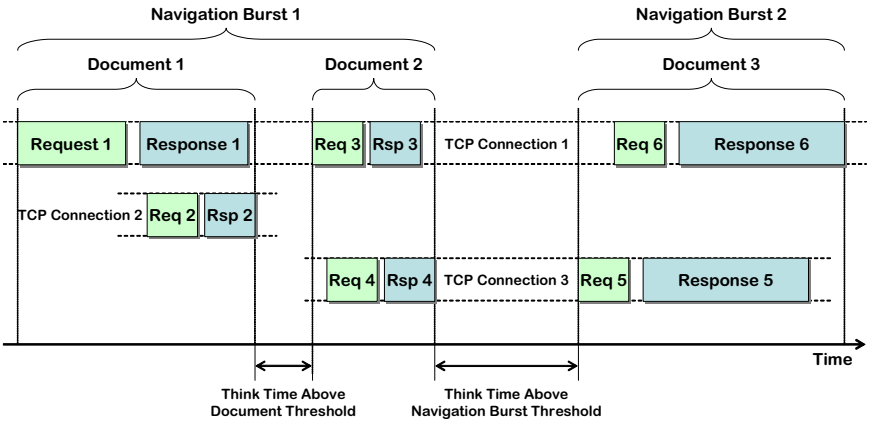


Figure 3. Navigation bursts group Web request/response pair separated by inactivity periods with a duration below some threshold value.

some or all of these persistent connections may be used to carry additional request/response pairs for the second document. In this case, documents do not provide a higher level aggregation of connections, but rather a different way of grouping responses.

By definition, Web documents could be closely related to user behavior. While the responses in a Web document start regardless of human choices, Web users select documents which they want to browse. However, as seen in the wavelet spectra in Figure 1, this is only partly true and the SiZer plot of document arrivals in Figure 2(b) supports this conclusion. The plot has less features compared to the response arrivals in (a). At the coarsest scale (the top row of the SiZer map), the estimated intensity increases on the left and decreases on the right. This is an artifact of the SiZer boundary adjustment caused by the "mirror image" approach (see Section 2 of [20]). However, aside from the boundary effect, the estimated intensity still has many oscillations, and the SiZer map confirms that they are statistically significant across all resolutions. Thus, the intensity is not anywhere near a constant. Combining the result of the wavelet spectra in Figure 1, we conclude that document arrivals have more Poisson properties than response arrivals, but still exhibits substantial LRD.

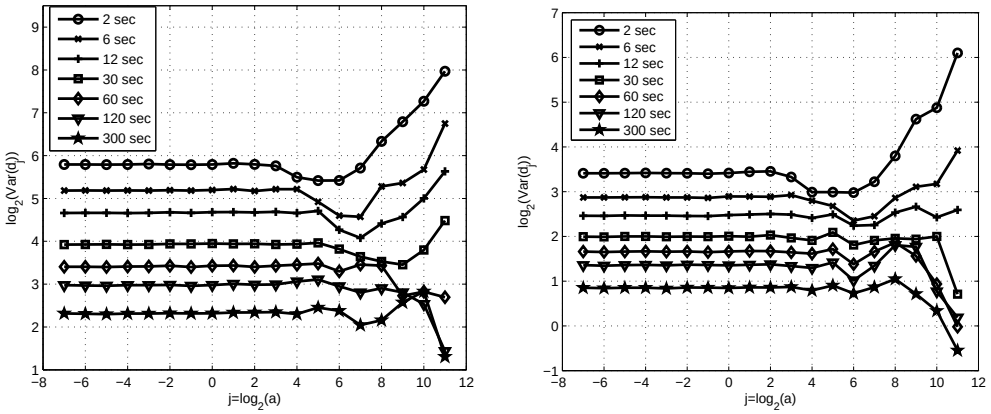
2.5. Clients

The final level of aggregation is a Web client, which corresponds to the activity of a single user during an entire trace. In our analysis, we group all of the documents downloaded by the same UNC IP address into a Web client. This makes it possible to extract the arrival process of clients from our traces, using the arrival time of the first document as the start time of the client. Intuitively, the client arrival process should exhibit clear Poissonity, since it is directly caused by human behavior.

Figure 2(c) depicts a SiZer plot of the client arrivals. The family of smooths shows a big decreasing trend at the beginning, which is artificially created by the definition of clients. Since our traces were collected during a four-hour time block, there are many clients who already started Web browsing before the trace collection. Beyond this starting region, the estimated intensity is mostly flat, which is consistent with a constant Poisson intensity. This can be confirmed by its SiZer map, which is located in the lower panel. It shows no features other than the big decreasing trend. Based on this SiZer plot and the wavelet spectra in Figure 1, the arrival process of clients is consistent with a homogeneous Poisson process. However, the client level has two undesirable characteristics, which motivated us to look for a finer level of aggregation consistent with Poissonity. Firstly, a strong boundary effect exists at the beginning of the collection period. Secondly, the sample size of the client level data is rather small. In the case of the 2001THU trace, only 17,295 clients were originally collected while 1,049,509 documents were collected during the same time block. If we remove clients which started before 1 PM, the dramatic decrease in sample size at the client level becomes even more substantial.

2.6. Filtering Problematic Clients

Having observed that document start times showed LRD scaling behavior against our initial intuition, we reexamined our data to verify that unusual clients (not associated with single-user Web browsing) are not behind our findings. This was a possibility, since our measurement heuristics can be confused by Web traffic that does not originate from a sin-



(a) 2001 Thursday

(b) 2001 Sunday

Figure 4. Wavelet spectra of navigation burst arrivals of (a) 2001THU and (b) 2001SUN at the thresholds of 2, 6, 12, 30, 60, 120 and 300 seconds. The plots suggest that a threshold between 12 and 30 seconds is the smallest which still gives a Poisson process.

gle human user browsing the Web. Problematic cases include traffic from time-shared machines, HTTP proxies, automated downloaders (*i.e.*, personal crawlers), SOAP (RPC over HTTP) and non-HTTP traffic in port 80. Given our knowledge of the UNC network, we did not believe that such phenomena was common enough to be significant in our datasets, but it was important to confirm this assumption.

Our analysis of "unusual clients" employed the following 8 criteria, where r is number of responses, τ is response inter-arrival, and d is duration:

(C1) Clients with $r > 3,000$.

(C2) Clients with a connection where $d > 2$ hours.

(C3) Clients with $r > 5$ responses, regularity index of responses above 0.8, and the median τ above 1 second.

(C4) Clients with $d > 3.5$ hours.

(C5) Clients with a maximum τ above 10,000 second.

(C6) Clients with a document where $d > 250$ seconds.

(C7) Clients with a number of connections above 3,000.

(C8) Clients with $d > 2$ hours, and very low think times.

Here *regularity index* is defined as follows. First, take the inter-arrival times within each document and calculate the median. Second, define the interval ($0.5\times$median, $1.5\times$median). Third, calculate the regularity index as the proportions of the response inter-arrival times that are covered by those intervals.

We systematically explored the impact of unusual clients by filtering out from our datasets any client satisfying some subset of the criteria. This filtering left 93.38%, 88.95%, and 82.31% of the clients for 2001THU, 2001SUN, and 2002SUN data, respectively. For the remaining clients, we reanalyzed document and client start times using plots of wavelet spectra and SiZer maps, and found little variation in the results. Our complete analysis is available online [13]. Therefore, it is unlikely that our conclusions are affected by unusual clients, or by inaccuracies of the measurement heuristics. One intuitive explanation is that our use of robust methods of statistical analysis, combined with a massive number of regular clients in our traces, makes our results very difficult to skew by unusual clients.

3. Analysis of Navigation Bursts

As discussed in Section 2, the arrival process of Web documents deviates significantly from a Poisson process, which suggests that it cannot be directly mapped to independent human behavior. The client arrival process is Poisson-like, but client level objects are too large. We are therefore led to consider a level of aggregation between document and client. Our choice is motivated by the observation that human browsing behavior usually alternates between two types of periods:

- A *navigation burst* period, in which a sequence of Web documents is downloaded with little inactivity (*i.e.*, short time intervals) between them. Users are often looking for some specific content, and in order to reach it, they have to navigate through several Web docu-

ments. This navigation burst creates a dependence between the arrivals of these documents that is inconsistent with Poisson arrivals.

- A *think time* period, in which the user reads or watches the content obtained at the end of the previous navigation burst period.

Navigation burst periods define a level of aggregation above documents but below clients, which we call the navigation burst level. Extracting the process of burst arrivals from our data sets requires the same technique used to extract the process of document arrivals. Think time analysis for extracting navigation burst information is illustrated in Figure 3. This definition is identical to our definition of document, but the inter-burst think time threshold is larger than the inter-document think time threshold. This implies that a navigation burst can be equivalently defined as a set of one or more documents separated by think times below the inter-burst think time. Obviously, it is crucial to find a think time threshold that can accurately extract navigation bursts from our data. The concept of navigation burst has not been studied in the past (at least in a similar context), so we have to carefully examine the question of an appropriate threshold.

3.1. Choosing a Think Time Threshold For Navigation Bursts

The threshold values on the think times we considered were 2, 6, 12, 30, 60, 100, 120, 300, 600, 1200, and 3600 seconds. For smaller threshold values, the process of navigation burst arrivals becomes closer to the arrival of documents. For larger threshold values, the process of navigation burst arrivals becomes closer to arrival of clients. The question is therefore to find the threshold at which a Poisson process becomes a reasonable model for navigation burst arrivals. We searched the range of threshold values on the gaps between two documents, which start to capture Poisson properties. When these values are too small, navigation burst arrivals will still have LRD just like document arrivals. If they are too large, navigation burst arrivals will suffer from severe boundary effects at the beginning and make the sample size small like client arrivals. We claim that the threshold values between 12 and 30 seconds seem to be where the Poissonity reveals itself, as shown below by several statistical analyses.

Since we analyzed four-hour traces, some navigation bursts start before the trace collection while some do not completely end when our collection is terminated. Therefore, in this section, we only use those "fully-captured" navigation bursts, which are defined as navigation bursts which only have fully-captured responses during the four-hour collection period.

Figure 4 depicts the wavelet spectra of the navigation burst arrivals for the 2001THU and 2001SUN traces with threshold values of 2, 6, 12, 30, 60, 120, and 300 seconds. (These thresholds are chosen to save space and still show the story clearly.) As one can see from the plots, the navigation bursts arrivals for small thresholds (2 and 6 seconds) have increasing trends at large scales, which suggest LRD. For the thresholds of 12 and 30 seconds, the spectra start to look like a Poisson spectrum although they are not quite flat at large scales. Therefore, we favor navigation bursts separated by 12- to 30-second think times.

As in earlier section, we used SiZer maps to further study the Poissonity of navigation burst arrivals. We first examine 2001THU in Figure 5. The SiZer map corresponding to the 6-second threshold is depicted in Figure 5(a), and it still shows complex trends like document arrivals in Figure 2(b). This suggests that Poisson processes are not suitable for the navigation burst arrivals under this small threshold.

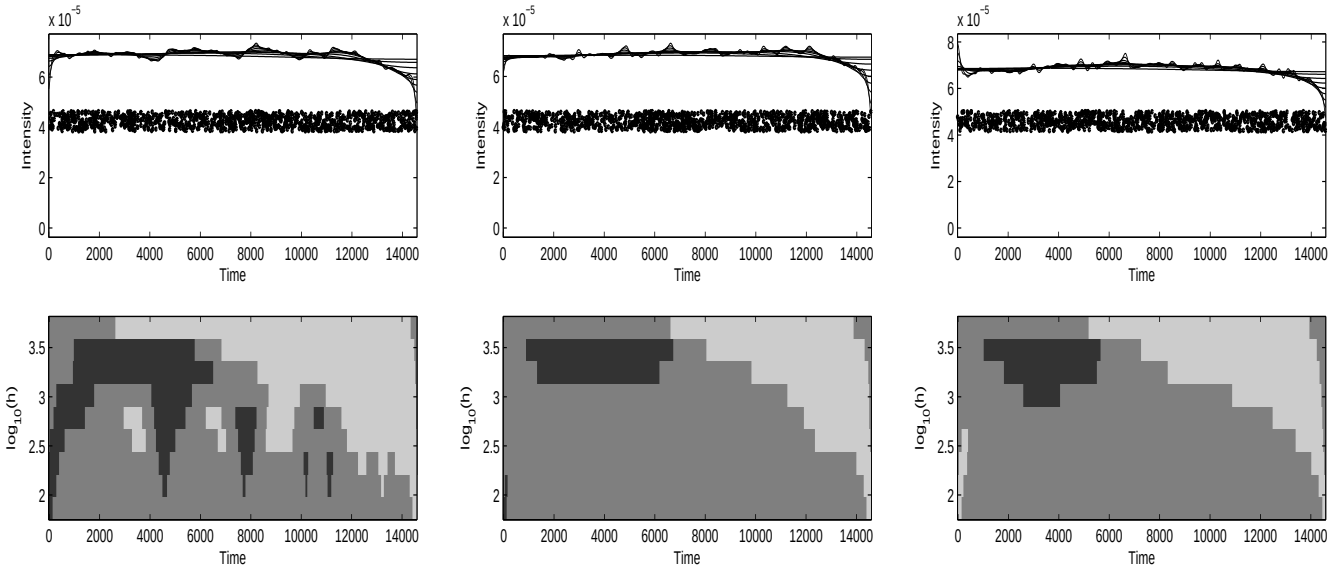
Figure 5(b) shows much less significant features compared to Figure 5(a). The top panel shows the kernel intensity estimates of the navigation burst arrivals with the threshold of 30 seconds. This plot shows that the estimated intensity looks like a constant function. The corresponding SiZer map shows significant increasing trends from $x = 2000$ to $x = 6000$ seconds. Except this trend, and the decreasing trend due to the boundary estimation problem of SiZer, the other features can be explained by natural variations and the intermediate gray colors on the SiZer map confirm this. Thus, one can conclude that the estimated intensity is close to a constant Poisson intensity.

Figure 5(c) shows the SiZer plot of the navigation burst arrivals with the threshold of 120 seconds. The SiZer plot looks similar to (b) except the decreasing trend at the beginning. This decreasing trend actually appears starting from the threshold of 100 seconds and becomes more serious as the threshold value increases. The reason is that as the threshold value increases, more and more navigation bursts start before the collection period. This is exactly the same boundary effect phenomenon which happens at the client level. These SiZer plots appear at [13].

Based on both the wavelet spectra and the SiZer plots, we claim that navigation bursts with the threshold values between 12 and 30 seconds are the regions where Poisson properties appear in the arrival processes.

3.2. Studying the Density of Navigation Bursts

In addition to these analyses, we develop an alternative way for validating our definition of navigation bursts using a think time threshold between 12 and 30 seconds. For this purpose, we studied the effect of different threshold on the set of navigation bursts. Our study relies on several con-



(a) 6 seconds (b) 30 seconds (c) 120 seconds

Figure 5. SiZer plots of navigation burst arrivals of 2001THU at the thresholds of (a) 6, (b) 30 and (c) 120 Seconds.

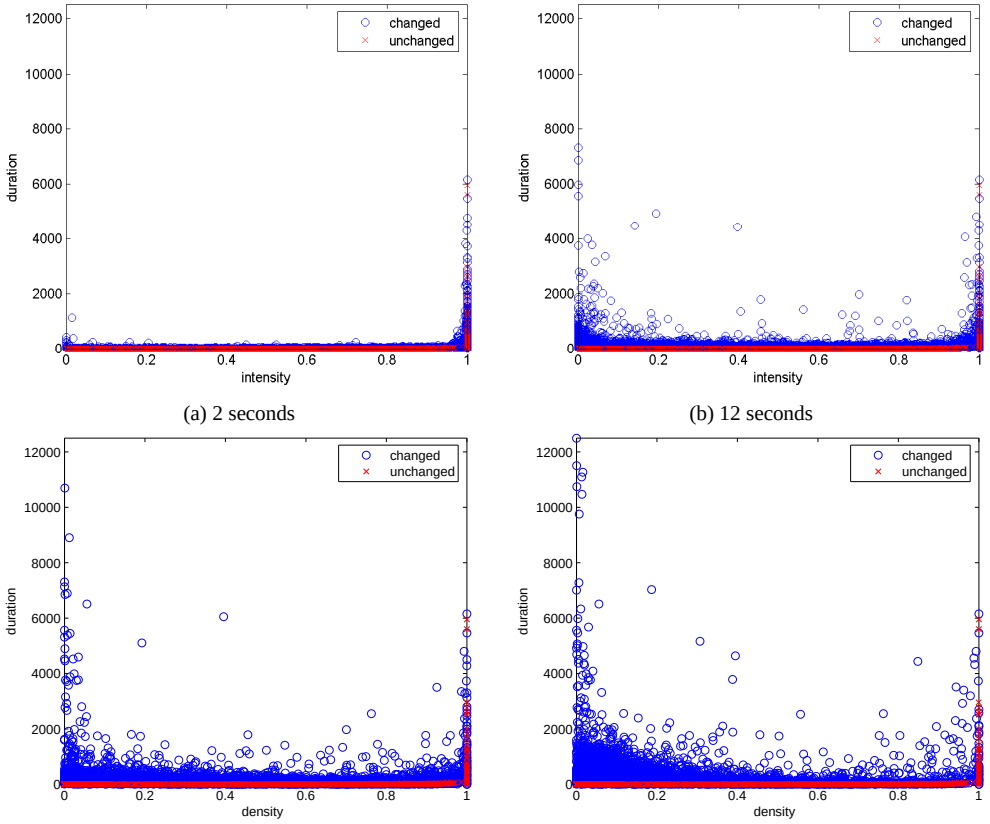
cepts: *Unchanged navigation bursts* are defined as navigation bursts whose start times and end times remain the same irrespective of the threshold. In contrast, *changed navigation bursts* are defined as navigation bursts whose start times and/or end times change (at least once) as the threshold value changes. Finally, the *density of a navigation burst* is defined as

$$\text{Burst Density} = 1 - \frac{\text{Total think time}}{\text{Duration}},$$

where *total think time* is the sum of the think times between the documents in the navigation burst, and *duration* is the total duration of the navigation burst, from the first data packet of the first response to the last data packet of the last response. By definition, navigation bursts with small threshold values are expected to have higher burst density, while navigation bursts with large thresholds are expected to have lower burst density. This is because navigation bursts get combined as the threshold value increases, which creates more think times within a navigation burst.

Figures 6(a)-(d) show the scatter plots of navigation burst durations versus burst densities of the 2001THU trace for the thresholds of 2, 12, 30, and 120 seconds, respectively. Unchanged bursts are marked as crosses and changed burst are marked as circles. As the threshold changes, only those changed bursts move. Note that, as the threshold increases, the circles (changed bursts) move from the bottom up because the durations of the changed bursts increase as well. These four plots show the way the intensity of the navigation bursts evolves as the burst threshold is modified. As one can see, for a small threshold (2 seconds), the changed bursts stack up around intensity 1, which correspond to the big cluster around intensity 1 in Figure 6(a). The fraction of high intensities decreases to wards zero (*i.e.*, moves to the left) as the threshold increases. As for the thresholds of 12 and 30 seconds, Figure 6(b)-(c) show a rough balance of the changed bursts between intensity 0 and 1, uncovering a clear *phase shift* for these threshold values. For the threshold of 120 seconds (Figure 6(d)), many changed bursts are from intensity 1, clustering around intensity 0. The plot for 300 seconds is shown at [13], and it has a big cluster around the intensity 0. The thresholds of 12 and 30 seconds seem to correspond to an intensity balance between 0 and 1. Thus, these two values can be understood as the points where the phase shifts from LRD (small thresholds, intensity 1) to Poisson (large thresholds, intensity 0).

Figure 7 shows this phase shift more clearly. We plot the fraction of changed bursts whose burst densities are greater than .95 (solid) and less than .4 (dashed) respectively against the threshold values for the 2001THU and the 2001SUN traces. The idea of this plot is to find the place where these lines experience sudden changes, *i.e.*, where a big shift happens. The plots reveal knees at both 12 and 30 seconds for both lines. This finding confirms that a phase shift happens around 12–30 seconds and this is another clear explanation of the starting point of the Poisson properties.



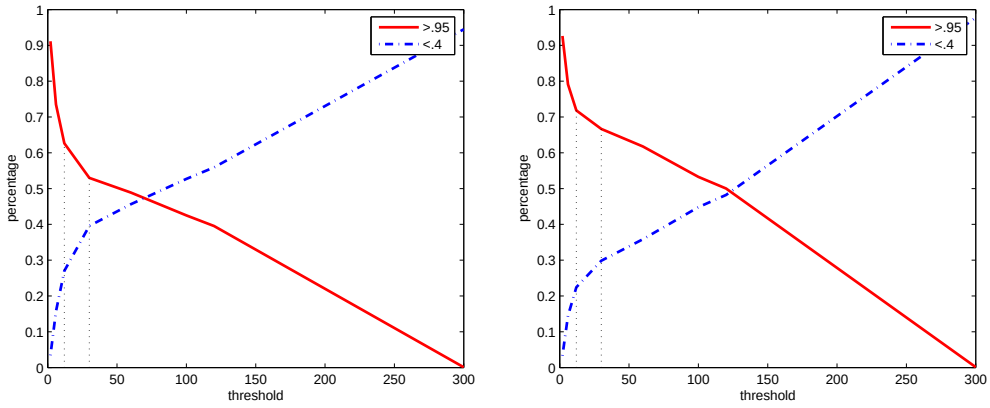
(a) 2 seconds

(b) 12 seconds

(c) 30 seconds

(d) 120 seconds

Figure 6. Scatter plots of navigation burst durations versus navigation burst densities of 2001THU for the four different thresholds. Changed navigation bursts are roughly balanced between the intensities 0 and 1 for 12 and 30 seconds.



(a) 2001 Thursday

(b) 2001 Sunday

Figure 7. Plots of % of changed navigation bursts with burst density $< .4$ and $> .95$. The sharp knees occur between 12 and 30 seconds highlighted by the two vertical dot lines.

The physical explanation is that Web users tend to search the Web by clicking several Web pages for a while, and then start to read articles of interest. Thus, after some amount of clicking time, burst arrivals are driven by independent user behavior which can be modeled as a Poisson process.

To complete the analysis, we also tried different versions of bursts by changing the definition of think time, *e.g.*, between request/response pairs rather than between responses. The results were similar for all definitions that we considered. The same analysis is applied to 2001 and 2002 Sunday mornings (between 8 AM and noon), and the results are consistent. A complete analysis with other statistical tools and different datasets is accessible at [13].

4. Conclusion

We studied various aggregation levels of Web traffic using several statistical tools and found the finest level of packet aggregation that exhibits Poisson properties. We show that the arrivals of Web responses, TCP connections and Web pages do not provide a satisfactory seed process. However, we found Poissonity in the arrivals of navigation bursts, which we defined as groups of Web documents separated by 12 to 30 seconds of idle time. This is consistent with common browsing behavior, which can be roughly divided into an active phase, where the user quickly follows links in the search for the desired content, and an inactive phase, where the user reads content more carefully.

We believe that our use of wavelet analysis and SiZer maps, together with the concept of burst density, provides a useful and robust methodology for uncovering Poissonity in the context of traffic modeling and other related problems. In a future project, we intend to develop a full model of browsing behavior whose foundation is the use of navigation bursts as the seed Poisson process. The model will capture the characteristics of packet, responses and document arrivals in a consistent and parsimonious manner.

References

[1] P. Abry and D. Veitch. Wavelet analysis of long–range dependent traffic. *IEEE Tran. on Information Theory*, 44:2–15, 1998.

[2] P. Barford and M. E. Crovella. Generating representative web workloads for network and server performance evaluation. In *Proc. of ACM SIGMETRICS*, pages 151–160, 1998.

[3] T. Berners-Lee, R. Fielding, and H. Frystyk. RFC 1945: Hypertext Transfer Protocol — HTTP/1.0, May 1996. Status: INFORMATIONAL.

[4] P. Chaudhuri and J. S. Marron. Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823, 1999.

[5] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. In *Proc. of ACM SIGMETRICS*, pages 160–169. ACM Press, 1996.

[6] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[7] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. RFC 2616: Hypertext transfer protocol – HTTP/1.1, 1999.

[8] F. Hernández-Campos. Multi-level HTTP data sets. `http://www-dirt.cs.unc.edu/multi-level-http`, 2005.

[9] N. Hohn, D. Veitch, and P. Abry. Cluster processes, a natural language for network traffic. *IEEE Tran. on Signal Processing*, 51:2229–2244, 2003.

[10] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Tran. on Networking*, 2(1):1–15, 1994.

[11] B. A. Mah. An empirical model of HTTP network traffic. In *Proc. of IEEE Infocom*, 1997.

[12] C. Nuzman, I. Saniee, W. Sweldens, and A. Weiss. A compound model for TCP connection arrivals for LAN and WAN applications. *Computer Networks*, 40(3):319–337, 2002.

[13] C. Park. Characterization of flow arrivals. `http://www-dirt.cs.unc.edu/semiexps`, 2005.

[14] C. Park, F. Hernández-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] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Tran. on Networking*, 3:226–244, 1995.

[16] J. Postel. RFC 793: Transmission control protocol, Sept. 1981.

[17] F. D. Smith, F. Hernández-Campos, K. Jeffay, and D. Ott. What TCP/IP protocol headers can tell us about the web. In *Proc. of ACM SIGMETRICS*, 2001.

[18] S. Stoev, M. Taqqu, C. Park, and J. S. Marron. On the wavelet spectrum diagnostic for hurst parameter estimation in the analysis of Internet traffic. *Computer Networks*, 48:423–445, 2005.

[19] D. Veitch. Matlab code for the wavelet spectrum. `http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html`, 2002.

[20] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

[21] W. Willinger and V. Paxson. Where mathematics meets the internet. *Notices of the American Mathematical Society*, 45(8):961–970, 1998.

[22] 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 Tran. on Networking*, 5(1):71–86, 1997.