IP Packet Generation: Statistical Models for TCP Start Times Based on Connection-Rate Superposition

William S. Cleveland
Statistics Research
Bell Labs, Murray, Hill, NJ
wsc@bell-labs.com

Dong Lin
Networked Computing Research
Bell Labs, Murray Hill, NJ
dong@bell-labs.com

Don X. Sun
Statistics Research
Bell Labs, Murray Hill, NJ
dxsun@bell-labs.com

ABSTRACT

TCP start times for HTTP are nonstationary. The nonstationarity occurs because the start times on a link, a point process, are a superposition of source traffic point processes, and the statistics of superposition changes as the number of superposed processes changes. The start time rate is a measure of the number of traffic sources. The univariate distribution of the inter-arrival times is approximately Weibull, and as the rate increases, the Weibull shape parameter goes to 1, an exponential distribution. The autocorrelation of the log inter-arrival times is described by a simple, two-parameter process: white noise plus a long-range persistent time series. As the rate increases, the variance of the persistent series tends to zero, so the log times tend to white noise. A parsimonious statistical model for log inter-arrivals accounts for the autocorrelation, the Weibull distribution, and the nonstationarity in the two with the rate. The model, whose purpose is to provide stochastic input to a network simulator, has the desirable property that the superposition point process is generated as a single stream. The parameters of the model are functions of the rate, so to generate start times, only the rate is specified. As the rate increases, the model tends to a Poisson process. These results arise from theoretical and empirical study based on the concept of connection-rate superposition. The theory is the mathematics of superposed point processes, and the empiricism is an analysis of 23 million TCP connections organized into 10704 blocks of approximately 15 minutes each.

1. MOTIVATION

Research on network traffic management heavily depends on simulations and laboratory experiments with synthetic traffic as input [19; 20]. Previous studies [24] on traffic modeling suggest that simple Poisson modeling of network traffic does not represent the characteristics of the actual aggregate traffic. Many traffic variables are long-range persistent: autocorrelations are positive and decay slowly [16; 17; 12; 9]. However, open-loop simulations, that is, without feedback, driven by stochastic modeling alone that models this dependence, tend to exaggerate the impact of persistence. Feedback-based TCP congestion control ameliorates to some degree the ill effects of persistence and responds well to increased network resources such as capacity and buffering [22; 23].

We are developing a closed-loop aggregate IP traffic simulation system using TCP congestion control. Our goal is to make the synthetic output traffic stochastically similar to that from the actual live wire of an Internet link. In our system, a TCP simulator is driven by stochastic inputs and puts out packet traffic that interacts with a network environment. The simulated TCP, which might spawn hundreds or thousands of TCP flows, responds to network congestion signaled via feedback packets. To mimic the behavior of TCP, we take the source code of an actual TCP implementation from the BSD kernel. Unlike the open-loop approaches which directly control the packet arrival times, packets are generated by TCP based on the processing of returning acknowledgment packets. The simulated network environment models propagation delays, link capacities, switching, routing, and congestion control inside the network. Packets traverse the simulated network and get dropped or arrive at the end host. This end host is also modeled by the TCP simulator.

The stochastic inputs in our simulation system include TCP connection start times for different applications, transferred file sizes, and the end-to-end connection propagation delays. The inputs are generated by statistical models that are developed based on mathematical theory and on empirical studies of packet header data collected on Internet wires.

This paper presents a study of start times under HTTP, both theoretical study and empirical study, which results in a statistical model that provides stochastic generation of starts for HTTP. Different protocols require different statistical models [24; 11], but the methods we employ to identify the HTTP model can be applied to other applications.

2. PREVIOUS RESULTS

TCP start times have been studied for a number of applications including HTTP, FTP, Telnet, and SMTP. Results have dealt with the marginal distributions of the inter-arrival times, autocorrelation, and cyclic patterns in the rates. As expected, the start-time rate has daily and weekly patterns because network usage has such patterns [24; 11]. However, it has been assumed in these papers that for one-hour intervals, the rate is stable. Inter-arrival times have been found to have a univariate distribution that is either exponential or has longer tails than the exponential; for HTTP, the distribution has been consistently reported to have longer tails [24; 21; 8; 11]. In the latter two papers, the distribution is found to be well approximated by the Weibull distribution with a shape parameter less than 1. The HTTP start times are long-range persistent with an estimate of the Hurst parameter in the vicinity of 0.75 [11; 10]. Feldman [11] explores the performance of an i.i.d. Weibull model and finds, even without the autocorrelation, that it does a better job of explaining connection admission blocking probabilities than does a nonstationary Poisson model.

Several papers contain discussions that provide an understanding of this behavior of HTTP connection starts [24; 1; 10; 11]. A single user clicks on Web links through time. For HTTP1.0, the dominant version of HTTP, a click results in an HTTP connection start for the linked file followed by connection starts for the embedded files. Browsers allow simultaneous transfer of files up to a maximum number. The times between the starts for these files can be much less than the times between clicks. So the end result is a burst of start times with small inter-arrival times and then a larger time until the next click. The magnitudes of the these small times for a single click tend to be related since they tend to be transfers from the same or related servers; for a close, lightly loaded server they will all tend to be small, but for a distant, heavily loaded server they will all tend to be larger. This behavior induces skewness in the marginal distribution of the inter-arrival times as well as positive autocorrelation, or persistence.

3. RESULTS OF THIS PAPER

3.1 Preliminaries

Let t_j for $j = 1$ to n be a sequence of HTTP inter-arrival times. The inverse of the mean of t_j is the rate ρ whose units we will take to be connection/sec, or c/s. The t_j can vary by several orders of magnitude, and small intervals are as important as large ones, so studying the data on a log scale is essential; $\log_2$, the log base 2, is more convenient than base 10 because we often need to consider variation that is a fractional power of 10.

Let $\ell_j = \log_2(t_j)$. The variation in ℓ_j can be decomposed into two components:

$$\ell_j = \ell_{1j} + \ell_{2j}, \tag{1}$$

where ℓ_{1j} is smooth daily and weekly variation, and ℓ_{2j} is the remaining variation, whose mean we take to be zero.

The results of this paper deal with the process ℓ_{2j}. We take time intervals for which the variation in ℓ_{1j} is negligible compared with the variation in ℓ_{2j}, so we effectively fix ℓ_{1j} to a specific value that becomes the mean of ℓ_j, and the inverse of this value is the rate ρ. We will study how the finite sample distributions of the ℓ_j change with ρ, but we do not develop a model for the statistical variation in ρ; however, in Section 8, we discuss how this can be approached, if needed, that is, if simulations are to be run over time intervals for which ℓ_{1j} changes by a nontrivial amount.

3.2 Nonstationarity

The ℓ_j are nonstationary. The nonstationarity occurs because the start times on a link, a point process, are a superposition of the source traffic point processes, and the statistical characteristics of superposition processes change as the number of superposed processes changes. The start time rate ρ is used as a measure of the number of traffic sources. As the number of sources changes, ρ changes, but the finite sample distributions of the ℓ_j change in ways much more complex than just the change in ρ. In particular, the univariate distribution and the autocorrelation function of the ℓ_j change.

3.3 Univariate Distributions

The univariate distribution of the inter-arrival times t_j is Weibull with shape parameter $\lambda(\rho)$ and scale parameter $\alpha(\rho)$, which depend on the rate ρ. From the properties of the Weibull [15],

$$\left(\frac{t_j}{\alpha(\rho)}\right)^{\lambda(\rho)} = u_j,$$

where u_j is a unit exponential, and

$$\mathrm{E}(t_j) = \rho^{-1} = \alpha(\rho)\Gamma(1 + \lambda^{-1}(\rho)). \tag{2}$$

A close approximation to the dependence of the shape parameter on ρ is

$$\log_2(\lambda(\rho)) = -1.963(1 + 1.275\rho^{0.3890})^{-1}. \tag{3}$$

As ρ increases, $\lambda(\rho)$ increases to 1, so at high rates, the inter-arrival distribution is close to exponential. From Equations 2 and 3, the dependence of the scale parameter on ρ is

$$\alpha(\rho) = \left[\rho\Gamma(1 + \lambda^{-1}(\rho))\right]^{-1}. \tag{4}$$

As ρ increases, $\rho\alpha(\rho)$ tends to 1.

On the log scale,

$$\ell_j = \log_2(t_j) = \lambda^{-1}(\rho)\log_2(u_j) + \log_2(\alpha(\rho)). \tag{5}$$

The ℓ_j have an extreme-value distribution with parameters $\lambda(\rho)$ [15] and $\alpha(\rho)$. The mean of ℓ_j is

$$\mu_\ell(\rho) = -\gamma\log_2(e)\lambda^{-1}(\rho) + \log_2(\alpha(\rho)) \tag{6}$$

where γ is Euler's constant (0.57722). The variance of ℓ_j is

$$\sigma^2_\ell(\rho) = \pi^2\log^2_2(e)/6\lambda^2(\rho). \tag{7}$$

3.4 Autocorrelation

A simple second-order model describes the second-moment properties of the ℓ_j in the sense that the power spectrum of the model closely fits the power spectrum of the ℓ_j estimated from the data by periodogram smoothing methods. The model is $s_j + n_j$. The n_j are a white noise series with variance $\sigma^2_n(\rho)$. The s_j are the long-range persistent series

$$(I - B)^{0.25}s_j = \epsilon_j + \epsilon_{j-1}, \tag{8}$$

where B is the backward shift operator, $Bs_j = s_{j-1}$, and ϵ_j is white noise with variance $\sigma^2_\epsilon(\rho)$ and is uncorrelated with n_j.

The two parameters of the second-order model are $\sigma_n^2(\rho)$ and $\sigma_\epsilon^2(\rho)$, which depend on ρ. The following describes this dependency. The variance of s_j, from [14], is

$$\sigma_s^2(\rho) = \frac{8\Gamma(1/2)}{3\Gamma^2(3/4)}\sigma_\epsilon^2(\rho). \quad (9)$$

Since n_j and ϵ_j are uncorrelated, n_j and s_j are uncorrelated, and

$$\sigma_\ell^2(\rho) = \sigma_s^2(\rho) + \sigma_n^2(\rho). \quad (10)$$

Let

$$\theta(\rho) = \sigma_n^2(\rho)/\sigma_\ell^2(\rho). \quad (11)$$

The functional form for $\theta(\rho)$ is well approximated by

$$\theta(\rho) = 1 - 2^{-1.2811-0.3150\log_2(\rho)}. \quad (12)$$

Thus

$$\sigma_n^2(\rho) = \theta(\rho)\sigma_\ell^2(\rho). \quad (13)$$

and

$$\sigma_\epsilon^2(\rho) = (1-\theta(\rho))\sigma_\ell^2(\rho)\frac{3\Gamma^2(3/4)}{8\Gamma(1/2)}. \quad (14)$$

As ρ gets large, $\sigma_\epsilon^2(\rho)$ and $\sigma_s^2(\rho)$ tend to zero, and $\sigma_n^2(\rho)$ tends to $\pi^2\log_2^2(e)/6$, so ℓ_j tends to white noise.

The autocorrelation of n_j at lag k, $a_n(k)$, is zero because n_j is white noise. The autocorrelation of s_j, from [14], is

$$a_s(k) = \frac{24-9k^{-2}}{16-9k^{-2}}\prod_{j=1}^{k}\frac{4j-3}{4j-1}.$$

$a_s(k)$ decreases with k and $a_s(1) = 5/7$. The autocorrelation of ℓ_j depends on ρ,

$$a_\ell(k,\rho) = (1-\theta(\rho))a_s(k).$$

If ρ increases from one value to a larger one, $a_\ell(k,\rho)$ drops by the same factor at each lag. For $\rho = 1$ c/s, 16 c/s, and 64 c/s, $a_\ell(1,\rho)$ is 0.29392, 0.12272, and 0.07930.

3.5 Statistical Generation Model

We built and validated a statistical model for generation of ℓ_j that balances simplicity and, to a good approximation, the reproduction of the extreme-value distribution, the autocorrelation, and the nonstationarity of these two aspects.

We begin by picking an overall process rate ρ for the generation, so this might be the average busy hour rate or any value on a curve that describes daily variation. This is all that needs to be chosen; the parameters of the statistical model are functions of ρ, as given above. (But see the discussion in Section 8 for calibration of ρ for a specific network.) There are two target specifications: (1) ℓ_j has an extreme-value distribution with parameters $\lambda(\rho)$ and $\alpha(\rho)$, and (2) ℓ_j has the autocorrelation function $a_\ell(k,\rho)$. The model for ℓ_j is

$$\ell_j = g_\rho(s_j + n_j), \quad (15)$$

where g_ρ, s_j, and n_j are defined next.

s_j is the series in Equation 8, but we add to it the property that it is a Gaussian process and the mean is zero; the s_j depend on the parameter $\sigma_\epsilon^2(\rho)$. n_j is the above white noise series with variance $\sigma_n^2(\rho)$, but we add the property that it is i.i.d. with an extreme-value distribution whose parameters are chosen so that the mean is $\mu_\ell(\rho)$ and the variance is $\sigma_n^2(\rho)$.

The $s_j + n_j$ have the target autocorrelation, but can be quite far from the target extreme-value distribution. We transform the $s_j + n_j$, producing ℓ_j that have exactly the target extreme-value distribution and very nearly the target autocorrelation. Let $W_\rho(u)$ be the cumulative distribution function of $s_j + n_j$. Let $Q_\rho(f)$ be the quantile of probability f of the target extreme-value distribution. Then the transformation that produces the target extreme value distribution is the composition function $Q_\rho(W_\rho(u))$. We found that this composition is well approximated by the function

$$g_\rho(u) = b_0(\rho) + b_1(\rho)u + b_2(\rho)u^2 \quad (16)$$

where

$$\begin{aligned} b_0(\rho) &= -e^{-0.7088-0.05857\rho} && (17)\\ b_1(\rho) &= 1 - e^{-1.6301-0.06399\rho} && (18)\\ b_2(\rho) &= -e^{-4.1896-0.06254\rho}. && (19) \end{aligned}$$

n_j always has an extreme-value distribution and its mean matches that of the target extreme-value distribution, but its variance is smaller than that of the target. However, as ρ gets large, the variance of s_j goes to zero, the variance of n_j tends to the target variance, the function g_ρ tends to the identity transformation, and ℓ_j becomes n_j.

This generation model ranges from a Poisson process at very high rates, to a long-range persistent process at low rates. Thus the model is significantly nonstationary.

The procedure for generating ℓ_j consists of the following computational steps: (1) $\lambda(\rho)$ from Equation 3. (2) $\alpha(\rho)$ from Equation 4. (3) $\sigma_\ell^2(\rho)$ from Equation 7. (4) $\theta(\rho)$ from Equation 12. (5) $\sigma_\epsilon^2(\rho)$ from Equation 14. (6) $\sigma_n^2(\rho)$ from Equation 13. (7) s_j from Equation 8 using methods in [14]. (8) μ_ℓ from Equation 6. (9) n_j as i.i.d. extreme-value distribution with mean μ_ℓ and variance $\sigma_n^2(\rho)$. (11) ℓ_j from Equations 15 to 19.

4. SUPERPOSITION, DATA, AND THEORY

In this section we introduce connection-rate superposition, and how we use it theoretically to study start-time point processes, and how we use it empirically to study start-time data. We also describe the data used in the empirical study.

A single user invoking HTTP on a network link creates a user TCP start-time point process. The start-time process for the link is the superposition of many such user processes. Furthermore, it is quite reasonable to suppose that the user processes are independent of one another within small time intervals provided there is not substantial local congestion and provided the users access many different Web servers; so we have a superposition of independent single-user processes.

We will not, however, build a user start-time model, and then create superposed user input by stochastic generation of many user streams; this approach has been taken by [18; 1]. Instead, we model the superposed process. This means that if there are 1000 users, in place of 1000 user streams we have just one, a superposition stream.

We model superposition based on the TCP connection rate, ρ. The basic assumption of this connection-rate superposition is that the statistical point process that generates the TCP starts for rate $\rho_k^* = k\rho_1^*$, is the k-fold superposition of k independent point processes with rate ρ_1^*. So long as ρ_1^* is

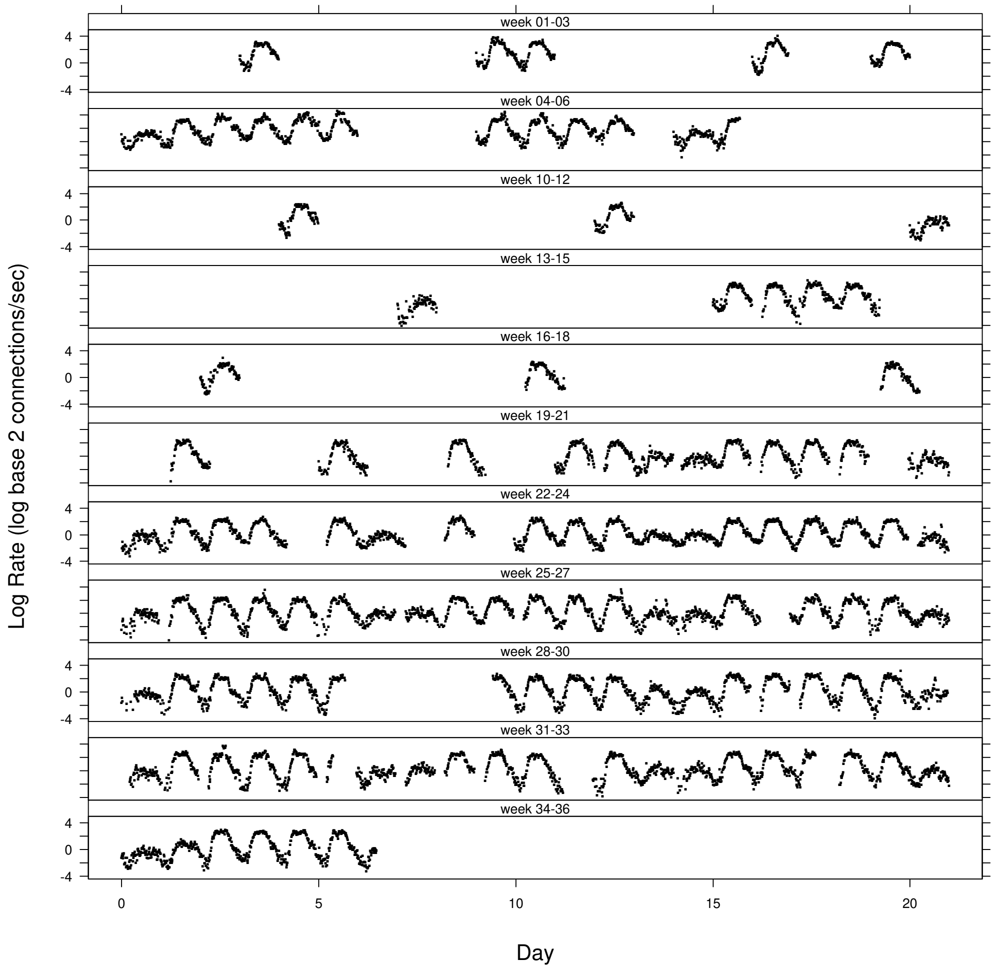

Figure 1: Log HTTP connection rate for 10704 blocks is graphed against time of occurrence of the block.

large enough to encompass one or more users, we can reasonably think of the k superposed processes as independent.

The empirical study in this paper is an analysis of 23 million HTTP start times measured at a single link that connects a Bell Labs network of about 3000 client hosts to outside servers that are widely distributed across the Internet. The hosts consist of PCs, multiuser Unix machines, and a Web cache. Inside the local campus network, congestion is low and round-trip times are negligible.

The study uses S-Net, a traffic collection and analysis system that begins with packet collection on a network link. Packet capture employs `tcpdump`, a 400 MHz PC, time-stamping based on GPS clock discipline, and attention to filter drops. The compressed header files are moved to a cluster of Linux PCs. Next, just TCP packets are processed; an algorithm organizes the header information by TCP connection flow, that is, by the source and destination ports and IP addresses. These flows are then processed to create data objects in S [3], a language and system for organizing, visualizing, and analyzing data. Carrying out the data analysis in S allows very rapid prototyping of new analysis tools tailored to the type of data. Daily monitoring commenced on 11/18/1998, and continues through the time of this conference, June 2000.

The data studied in this paper consist of TCP/IP packet headers for TCP connections under HTTP for the period 11/18/98 to 7/10/99. On 12/20/98, a Web cache was installed that at first served the whole network but on 1/9/99 was reduced to just one third of the hosts. We used data just from hosts other than the cache because it was under experimentation during this period. So there are gaps in our data resulting from the cache, and also from periods when the monitor was down.

We organized the data into 15-minute blocks. Each block consists of the TCP connections whose SYN packets arrived during the block. We want the block length to be as large as possible subject to the constraint that the smooth daily and weekly variation, the component ℓ_{1j}, is nearly constant. Our study of the data led to the conclusion that 15 minutes is an appropriate length. Not every block from 11/18/98 to 7/10/99 appears due to idleness and monitor down time. And, we eliminated blocks with less than 95% of the full 15 minutes, blocks with fewer than 50 flows, and connections from certain hosts that developed problems. The final result is data on 23,008,664 TCP connections organized into 10704 blocks. The number of flows in the blocks ranges from 52 to 22470; the median is 1149, the upper quartile is 3461, and the lower quartile is 548. We assume that the point process of start times is stationary within a block b.

Let t_{bj} for $j = 1$ to n_b be the inter-arrival times in seconds

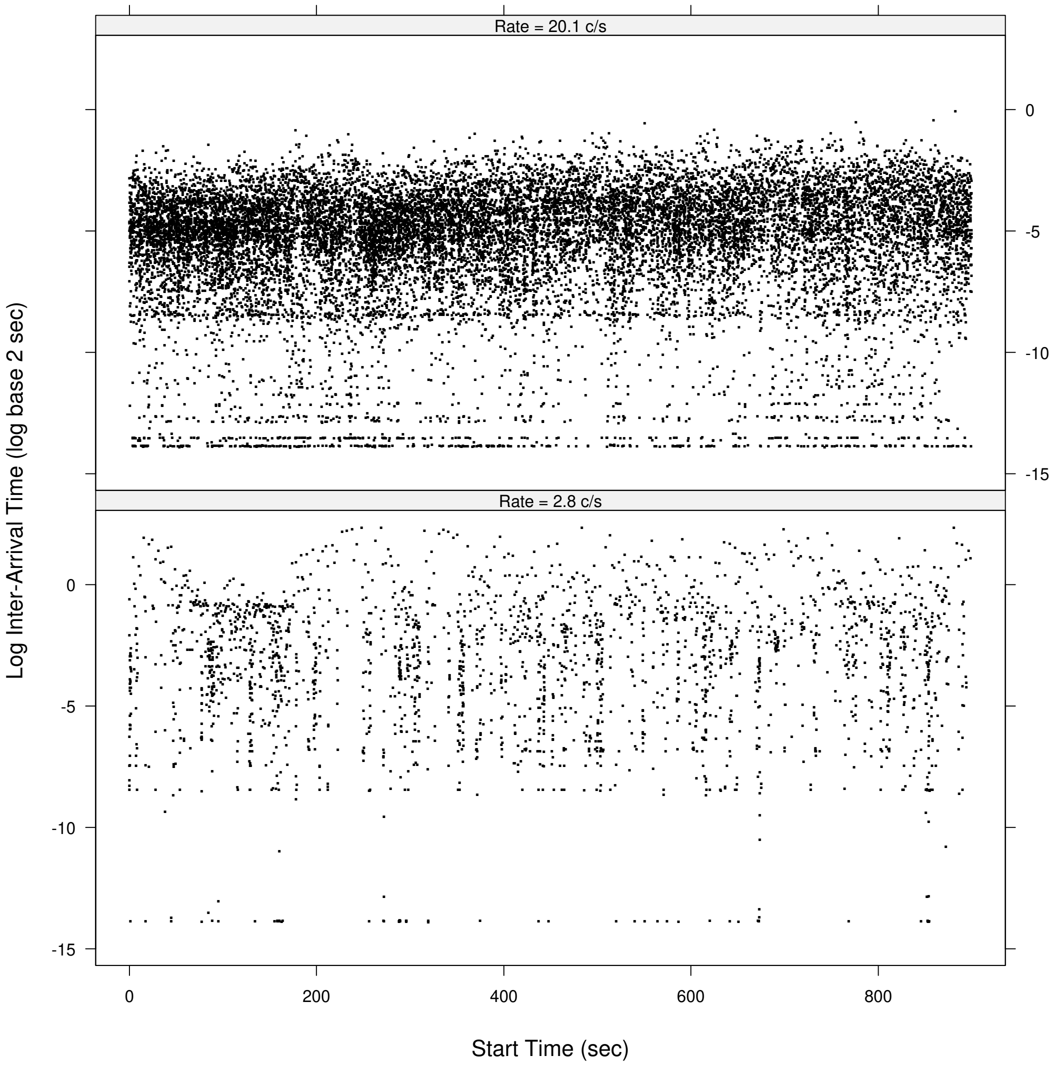

Figure 2: Log inter-arrival time is graphed against start time for two blocks of HTTP starts. The connection rate for each panel is shown in the strip label at the top of the panel.

and let $\ell_{bj} = \log_2(t_{bj})$. Let $\hat{\mu}_b$ be the sample mean of this distribution. Then the sample rate is $\hat{\rho}_b = \hat{\mu}_b^{-1}$. Figure 1 plots $\log_2(\hat{\rho}_b)$ against the time of block b. The major cycles in the data are the daily variation with reduced peaks on non-workdays. The $\hat{\rho}_b$ vary from $2^{-4.06} = 0.060$ c/s to $2^{4.64} = 25.0$ c/s.

Our overall strategy, invoking connection-rate superposition, is to study the statistical properties of the ℓ_{bj} in each block and see how these properties change with the sample rate $\hat{\rho}_b$. Figure 2 reveals changing properties. The bottom panel is an *inter-arrival plot* or *i-a plot* of the 2515 start times for one block with $\hat{\rho}_b = 2.80$ c/s. On the plot, the jth log inter-arrival time ℓ_{bj} is plotted against the time at the beginning of the interval for $j = 1$ to n_b. The log on the vertical scale is vital because, as we stated earlier, inter-arrival times can vary by several orders of magnitude. The horizontal scale, however, conveys arrivals and inter-arrivals on the original scale. The top panel of Figure 2 is an i-a plot for another 15 minute block on the same day. The sample connection rate, $\hat{\rho}_b = 20.1$ c/s, is about 7 times greater than that for the first block.

Both panels show discreteness on the vertical scale, many nearly equal values of values of ℓ_{bj} piling up at several locations such as $-14 \log_2$ sec. This is a network effect, a small delay; each accumulation point is the $\log_2$ of the time it takes to process a packet in the network. For example, suppose two SYN packets are back to back, which happens a small fraction of the time. They arrive on the wire, the first is time-stamped and then is read by the PC monitor card. Then the second is time-stamped, so the inter-arrival time is the time it takes to read the first packet.

In the bottom panel of Figure 2, the data form distinct vertical bands with ℓ_{bj} taking values in the middle of the distribution of values. These are bursts of connections caused by single clicks of individual users. The larger values of ℓ_{bj} on the plot tend to be quiescent times until the click of some user occurs. In the top panel of Figure 2, the bursty behavior at the small time scales has disappeared. Because the rate is much higher, the SYNs of more users intermingle, and the behavior of individual users is broken up. The plot shows another difference between the two panels; the variance of the ℓ_{bk} at the lower rate is greater.

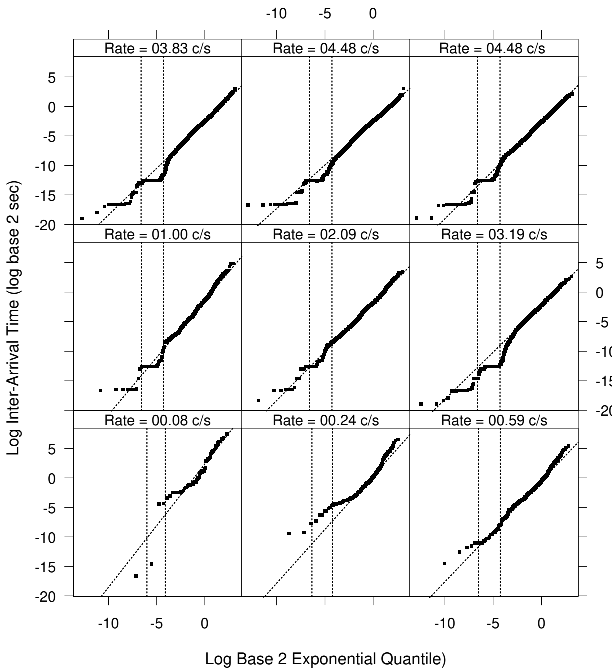

Figure 3: Quantiles of log inter-arrival times are graphed against log Weibull quantiles. The vertical lines show the 1% and 5% quantiles. The oblique line is drawn through the two quartile points.

The theoretical study also invokes connection-rate superposition. If we model the start-time point process for a base rate ρ_1^*, then we can derive mathematically the characteristics of the point process when the rate is $\rho_1^* = k\rho_1^*$, where k is a positive integer, by invoking the mathematical theory of point process superposition or by running simulations when the math is not tractable.

The empirical and theoretical study interact in the following way. The empirical study results in a characterization of the statistical properties of the ℓ_j and how these properties change with ρ. We use these characterizations to specify an empirical model for the base rate, ρ_1^*, and then we derive the statistical properties for rates $\rho_k^* = k\rho_1^*$. Finally, we compare the empirical results for sample rate $\hat{\rho}_b$ with the theoretical results at that rate. This process provides a much more powerful method of investigation of the statistical properties than we would have with just theory or just empiricism alone. With the statistical characteristics in place, we develop a statistical model for the generation of synthetic start times that reproduce the characteristics.

5. INTER-ARRIVAL DISTRIBUTION

In this section we study the distribution of the inter-arrival times empirically by analyzing the data described in Section 4. We use a data visualization tool to reveal the structure of the distribution. We relate changes in the distribution to the sample rate $\hat{\rho}_b$. Then, using mathematical results for the superposition of renewal processes, we build a model for the distribution that depends on the rate ρ, and validate this model from the data.

5.1 Empirical Study

For each of the 10704 blocks of inter-arrival times we used a data visualization tool, a quantile plot [4], to study the empirical distribution of ℓ_{bj} for $j = 1$ to n_b, the log inter-arrival times for block b. Trellis display, a framework for multipanel data display, made the task of displaying 10704 plots relatively easy [2]. The goal is to determine whether the empirical distribution is well approximated by some extreme-value distribution with parameters λ_b and α_b. The quantile plot is effective because it shows all of the data and allows us to study the approximation in detail across the entire range of the data.

Let $\ell_{b(j)}$ for $j = 1$ to n_b be the values of ℓ_{bj}, ordered from smallest to largest. The empirical quantiles are defined by taking $\ell_{b(j)}$ to be the quantile with (empirical) probability $f_j = (j - 0.5)/n_b$; approximately a fraction f_j of the data are less than or equal to $\ell_{b(j)}$. Let $H(u)$ be the cumulative distribution function of the extreme-value distribution with the two parameters equal to 1. Let h_j be the quantile of order f_j of H; this means that $f_j = H(h_j)$. On the quantile plot, $\ell_{b(j)}$ is plotted against h_j. From Equation 5, the quantile of order f_j of the extreme-value distribution with parameters λ_b and α_b is $h_j/\lambda_b + \log_2(\alpha_b)$. If the pattern on the quantile plot is close to a line, then the empirical distribution of the ℓ_{bj} is well approximated by the extreme-value distribution; the slope of the pattern estimates λ_b^{-1} and the intercept estimates $\log_2(\alpha_b)$.

Figure 3 shows extreme-value quantile plots for 9 blocks from one day. The strip label at the top of each panel shows the sample rate $\hat{\rho}_b$. The oblique line on the plot is drawn through the upper and lower quartile points. The two vertical lines on the plot are drawn at the 0.01 and 0.05 quantiles. The pattern of the points on the plot follows the quartile line quite well. There are deviations at the low end of the distribution caused by the discreteness discussed in Section 4. This a network artifact, not a reflection of the true start times of the client connections; the artifact affects only a small fraction of the data, never more than 5%, as the vertical lines in Figure 3 show.

We use the maximum likelihood estimate $\hat{\lambda}_b$ to estimate λ_b. Figure 4 graphs $\log_2(\hat{\lambda}_b)$ against $\log_2(\hat{\rho}_b)$. Under an assumption of independence of the t_{bj}, the variance of $\log_2(\hat{\lambda}_b)$ does not depend on λ_b, and is asymptotically equal to $1.2654/n_b$ [15]. $\hat{\rho}_b$ is approximately n_b divided by the block length, so $\hat{\rho}_b$ is approximately proportional to the inverse of the variance. Thus, in Figure 4, when $\hat{\rho}_b$ increase by a factor of, say, 2, the variance decreases by a factor of approximately 2.

The smooth curve on the plot was fitted by loess, a nonparametric regression curve estimator [5]; the loess smoothing parameter is 0.4, and the fitting is robust and locally linear. The curve was evaluated at 50 equally-spaced points from the smallest $\hat{\rho}_b$ to the largest, and is plotted by connecting successive points by line segments.

Figure 4 shows the strong dependence of $\hat{\lambda}_b$ on $\hat{\rho}_b$. For the smallest sample rate, $2^{-4.06} c/s = 0.060$ c/s, the loess fit gives $\lambda_b = 2^{-1.37} = 0.39$. For the largest sample rate, $2^{4.65} c/s = 25.0$ c/s, the loess fit gives $\lambda_b = 2^{-0.21} = 0.86$. This is a major difference; Weibull distributions with these two values of the shape parameter differ markedly from one another. Thus the nonstationarity in the HTTP start process is substantial.

In Figure 4 there are a very small number of values of $\hat{\lambda}_b$ that are large for their particular values of $\hat{\rho}_b$. The errant

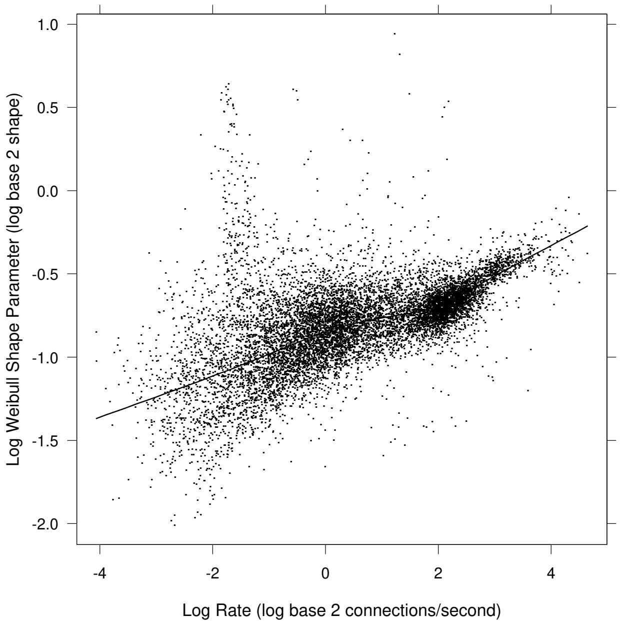

Figure 4: The log of the maximum likelihood estimate of the Weibull shape parameter is graphed against the log sample connection rate for the 10704 blocks. The smooth curve on the plot was fitted by loess, a regression smoothing method.

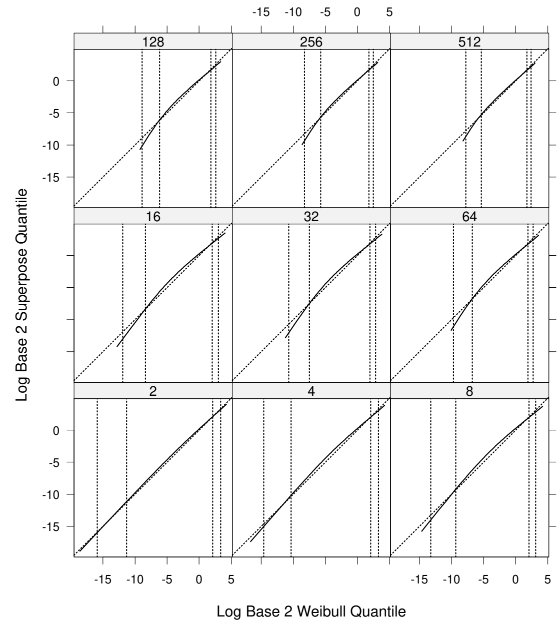

Figure 5: Log superposition quantiles are graphed against the log best-approximating Weibull quantiles for 9 values of k, shown in the strip labels at the tops of the panels. The vertical lines are drawn at the 0.01, 0.05, 0.95, and 0.99 quantiles.

values occur in the range where $\hat{\rho}_b$ is between -2 and 2. They are caused by clients opening large numbers of connections at equally-spaced intervals; this tends to push the empirical distribution of the t_{bj} toward the uniform, which increases the value of $\hat{\lambda}_b$. We do not see this effect for very small or very large rates $\hat{\rho}_b$. If one of the sources that generates these connections begins, the rate is increased above a minimum one. If other HTTP traffic has a high enough rate, the regular intervals are broken up sufficiently that they do not appreciably affect the distribution. We ignore this traffic in our modeling since it is atypical. Note that leaving the values in our data does not adversely affect the loess estimate, which is robust.

5.2 Theoretical Study

Section 6 will show that the start-time point process for low rates is correlated. But for the purpose of theoretically investigating the univariate distribution of the inter-arrival times, we will suppose they are independent, so the start times follow a renewal process. The assumption is reasonable for our purpose, because the magnitude of the autocorrelation is small, and the assumption succeeds in that it yields good predictions for the empirical data just analyzed.

Let ρ_1^* be a low base rate. Let s_1 be a random variable whose distribution is that of the inter-arrival times when the rate is ρ_1^*. The mean of s_1 is $1/\rho_1^*$. Suppose we superpose k start processes with rate ρ_1^*. Let s_k be a random variable whose distribution is that of the inter-arrival times of the superposition process. The rate for s_k is $\rho_k^* = k\rho_1^*$ and the mean is $1/\rho_k^*$. Now consider s_k divided by its mean, and to help curb the amount of notation, let us switch notational meaning and denote this normalized variable by s_k. Let $S_k(s)$ be the distribution function of this new s_k, and let $\bar{S}_k(s) = 1 - S_k(s)$ be the survival function. From standard results in the superposition of renewal processes [7],

$$\bar{S}_k(s) = \bar{S}_1(s/\rho_k^*)\left(\int_{s/\rho_k^*}^{\infty} \rho_1^* \bar{S}_1(u)du\right)^{k-1}. \qquad (20)$$

We will take the base rate to be $\rho_1^* = 0.060$ c/s, the minimum of the sample block rates in the empirical study. Because of the results of the empirical study, we will take the distribution of s_1 to be Weibull. We estimate the shape parameter of this distribution by a procedure that will be explained later; the resulting value is $\lambda_1^* = 0.40$. When a Weibull random variable with parameters λ and α is normalized by dividing by its mean, the result is a Weibull with shape λ and scale $\Gamma(1 + 1/\lambda)$. Thus the scale parameter for s_1 is $\alpha_1^* = \Gamma(1 + 1/\lambda_1^*)$.

With the univariate distribution of s_1 specified, we can use Equation 20 to derive the distribution of all s_k. We will do this for $k = 2$ to 416. The maximum rate, $\rho_{416}^* = 25.0$ c/s, is close to the maximum sample block rate in the empirical study. Let $G(z;\tau)$ be the cumulative distribution function of the gamma distribution with shape τ and scale 1. Then

$$\bar{S}_k(s) = e^{-\zeta_k(s)}\left(1 - G(\zeta_k(s); 1/\lambda_1^*)\right),$$

where

$$\zeta_k(s) = \left(\frac{s\,\Gamma(1 + 1/\lambda_1^*)}{k}\right)^{\lambda_1^*}.$$

$S_k(s)$ is not a Weibull distribution for $k > 1$, but we will check to see if it is well approximated by a Weibull with a mean of 1, shape λ_k^*, and scale $\Gamma(1+1/\lambda_k^*)$. We would expect this to be the case because the empirical study showed that the empirical distribution of the inter-arrival times is well approximated by the Weibull.

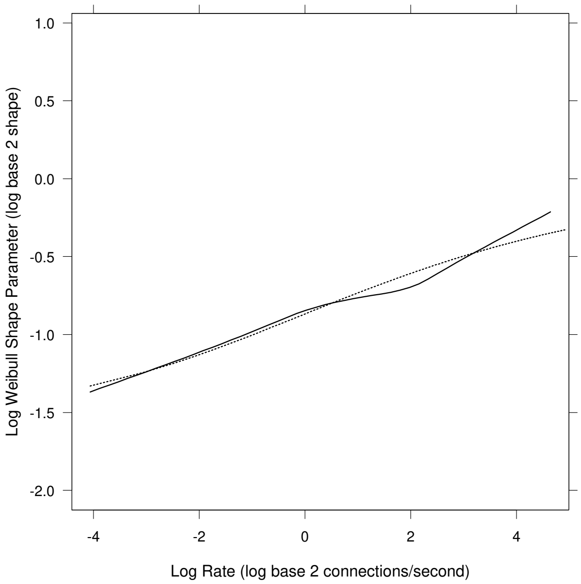

Figure 6: The solid curve is the loess fit displayed in Figure 4. The dashed curve is a plot of log shape against log connection rate for the Weibull that best approximates the superposition distribution at that rate. The scales have been chosen to match those of Figure 4 to enhance comparison.

We will base the approximation on extreme-value quantiles, just as we used these quantiles to assess the goodness of the approximation of the extreme-value distribution to the empirical distribution of ℓ_{bj}. We proceed in order from $k = 2$ to 416. For each k, we start with the best approximating normalized Weibull for $k-1$, which has parameters λ^*_{k-1} and $\Gamma(1+\lambda^*_{k-1})$. For $k = 2$, the starting distribution has shape λ^*_1. Let $f_i = (i - 0.5)/1000$ for $i = 1$ to 1000. Let $w_{(i)}$ be the quantile of probability f_i of the starting Weibull distribution. Let $g_i = S_k(w_{(i)})$. The $w_{(i)}$ are quantiles with probabilities g_i of $S_k(s)$. Suppose the $w_{(i)}$ are approximately Weibull quantiles with probabilities g_i, shape λ^*_k, and scale $\Gamma(1+1/\lambda^*_k)$. Then

$$\log(w_{(i)}) \approx \log(-\log(1-g_i))/\lambda^*_k - \log(\Gamma(1+1/\lambda^*_k)). \quad (21)$$

A best approximating value of λ^*_k is found by least squares fitting of the left side of approximate Equation 21 to the right side.

The resulting approximating Weibull quantiles provide a good fit to the superposition quantiles for all k. This is illustrated in Figure 5. For 9 of the above values of k, we plot $\log_2$ quantiles of the best approximating Weibull against the $\log_2$ quantiles of the superposition distribution. Figure 6 graphs the loess curve of Figure 4 (solid curve) and graphs $\log_2(\lambda^*_k)$ against $\log_2(\rho_k)$ (dashed curve). Clearly the superposition curve is in agreement with the pattern in the data.

We use the following optimization method to find λ^*_1. We select a trial estimate, and then compute λ^*_k for $k = 2$ to 416 by computing the superposition distributions and then finding the best approximating Weibulls. Thus we have λ^*_k for the rates ρ^*_k. We next compute shape parameters for all sample rates $\hat{\rho}_b$ for $b = 1$ to 10704 by linearly interpolating the 416 (ρ^*_k, λ^*_k) pairs in the rate. Suppose the resulting

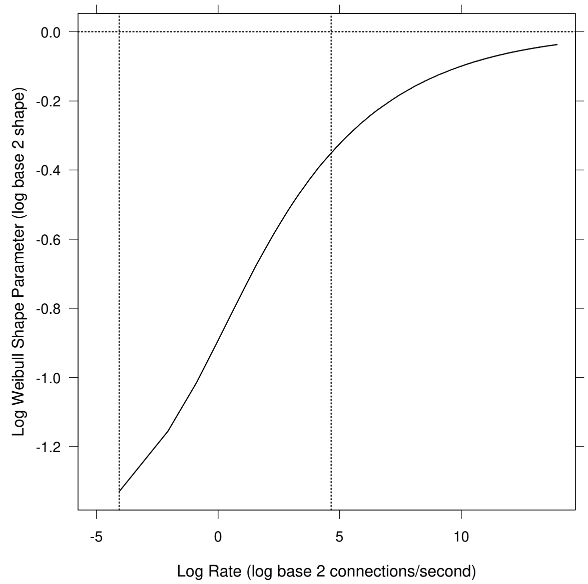

Figure 7: The log of the shape parameter of the Weibull that best approximates the superposition distribution at rate ρ is graphed against log ρ. The two vertical lines show the minimum and maximum sample block rates.

values are $\tilde{\lambda}_b$. The optimization criterion for the trial value of λ^*_1 is

$$\sum_{b=1}^{10704} \omega_b(\log(\hat{\lambda}_b) - \log(\tilde{\lambda}_b))^2,$$

where, as defined above, $\hat{\lambda}_b$ is the maximum likelihood estimate of λ_b. The weights ω_b are the product of the robustness weights from the loess fit and the variances of $\log_2(\hat{\lambda}_b)$, $1.2654 n_b^{-1}$, under an assumption of independent inter-arrival times. We select λ^*_1 to minimize this criterion; the resulting value is 0.40.

5.3 Discussion

The empirical and theoretical results are in close agreement. The univariate distribution of the HTTP start time inter-arrivals is well approximated by the Weibull. Our overall framework, connection-based superposition, has been validated in the sense that the theoretical results to which it leads fit the empirical results.

Because the results are grounded on theory, with the empirical study providing the validation, it is reasonable to apply the theory for rates beyond those of our data. Of course, this needs to be checked with data from a high-throughput network link, but packet capture is not readily available at the highest line speeds now in use. Figure 7 graphs $\log_2(\lambda^*_k)$ against $\log_2(\rho^*_k)$, where ρ^*_k is now $k\rho^*_1$ for $k = 1, 2^2, 3^2, \ldots, 524^2$. The two vertical lines show the maximum and minimum values $\log_2(\hat{\rho}_b)$; we have gone out beyond the data in our prediction by an amount slightly bigger than the range of the data. At 2^{14} c/s, the Weibull shape is 0.97, which is very close to exponential. Equation 3 results from least-squares fitting to these 524 points; the resulting fit provides an excellent approximation.

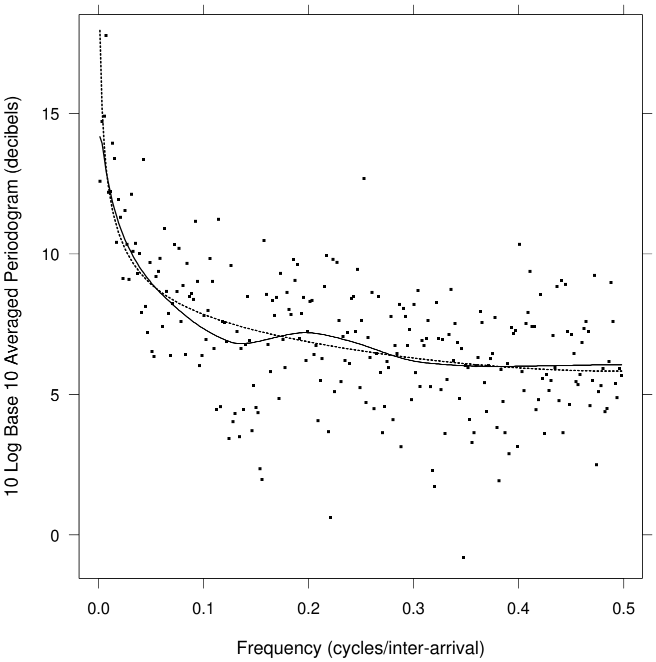

Figure 8: The log averaged periodogram is graphed against frequency for the data in the bottom panel of Figure 2. The solid curve is the ATS estimate of the log power spectrum and the dashed curve is an estimate from the second-order model.

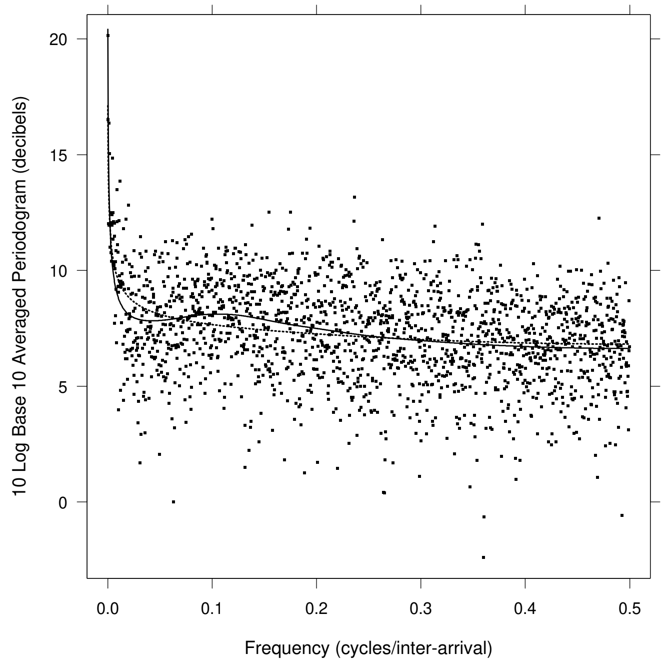

Figure 9: The display method of Figure 8 is used for the data in the top panel of Figure 2.

6. AUTOCORRELATION

In this section we study the second-order properties of the ℓ_{bj} empirically by estimating their power spectrum using periodogram smoothing methods, and characterize the dependence of the spectrum on $\hat{\rho}_b$, the sample rates. From this we identify, fit, and validate the second-order model presented in Section 3.4. Then, we investigate the second-order properties theoretically through simulation, studying the results by the same mechanism used to study the empirical data, estimating the power spectrum.

6.1 Empirical Study

We estimate the power spectrum for 500 of the 10704 blocks of times, randomly selected but constrained by the requirement that the selected $\hat{\rho}_b$ rates range from the lowest to the highest values of all of the $\hat{\rho}_b$. For convenience of notation, we take the selected blocks to be denoted $\hat{\rho}_b$ for $b = 1$ to 500. Let $p_b(f)$ be the block b power spectrum, that is, the power spectrum of the ℓ_{bj} for $j = 1$ to n_b.

Our first step is to estimate $p_b(f)$ by a nonparametric periodogram smoothing procedure, a method that provides a very flexible estimate, because it is constrained only by local smoothness. The purpose is to use the estimates for the 500 blocks to identify a second-order model and study how it changes with the sample block rates $\hat{\rho}_b$ for $b = 1$ to 500. The estimation is based on ATS methods [6]: (1) subtract the sample mean of the data, and compute the periodogram $I_b(f)$ at the Fourier frequencies $f_{bk} = k/n_b$, for $0 < f_{bk} \leq 0.5$; (2) average the periodogram in non-overlapping blocks of size 5 (dropping the last block if there are fewer than 5 in the average), and average the frequencies in the same way to form $\bar{I}_b(\bar{f}_{bk})$ at m_b equally-spaced frequencies $\bar{f}_{bk}$, for $k = 1$ to m_b; (3) take log base 10 and multiply by 10 (so that units are in decibels), yielding $z_b(\bar{f}_{bk}) = 10\log_{10}(\bar{I}_b(\bar{f}_{bk});$ (4) smooth the $z_b(\bar{f}_{bk})$ using loess (see Section 5) to form the ATS log power spectrum estimate $10\log_{10}(\hat{p}_b(\bar{f}_{bk}))$. The details of the loess smoothing are given later.

Figures 8 and 9 plot $z_b(\bar{f}_{bk})$ against $\bar{f}_{bk}$ for the two blocks of ℓ_{bj} graphed in Figure 2. The solid curve in each figure is $10\log_{10}(\hat{p}_b(\bar{f}_{bk}))$. The sample rate for Figure 8 is low, $\hat{\rho}_b = 2.8$ c/s. The sample rate for Figure 9 is higher, $\hat{\rho}_b = 20.1$ c/s. By studying such power spectrum plots we identified the model described in Section 3.4. The power spectrum of this model is

$$p_b(f) = \sigma^2_{b\epsilon}\frac{|1 + e^{2\pi i f|}|^2}{|1 - e^{2\pi i f}|^{2d_b}} + \sigma^2_{bn}. \qquad (22)$$

The parameters of the model are $\sigma^2_{b\epsilon}$, σ^2_{bn}, and d_b; they are estimated by fitting $p_b(f)$ to the $z_b(\bar{f}_{bk})$; the details of the fitting are given later. Let $\hat{\sigma}^2_{b\epsilon}$, $\hat{\sigma}^2_{bn}$, and $\hat{d}_b$ be the estimates. We found that $\hat{d}_b$ does not depend on $\hat{\rho}_b$; the median of $\hat{d}_b$ is 0.26, so we take $d_b = 0.25$ in the model (unable to resist the increased esthetic value in subsequent formulas) and re-estimate the remaining two parameters. The resulting model estimate of the power spectrum, $\tilde{p}_b(f)$, is Equation 22 with $d_b = 0.25$ and the other two parameters replaced by their estimates. $10\log_{10}(\tilde{p}_b(f))$ is graphed in Figures 8 and 9 by the dashed curves.

The second-order model provides an excellent fit in the sense that the power spectrum estimate from the model is quite close to the smoothing estimate from loess for most of the 500 blocks; for example, the two are close in Figures 8 and 9. There is, in fact, a minor, but common departure of the loess estimate from the model estimate, a peak with a period that varied between 10 to 20 inter-arrivals; such peaks occur in Figures 8 and 9. It is possible that this is caused by bursts of arrivals due to embedded files, with the

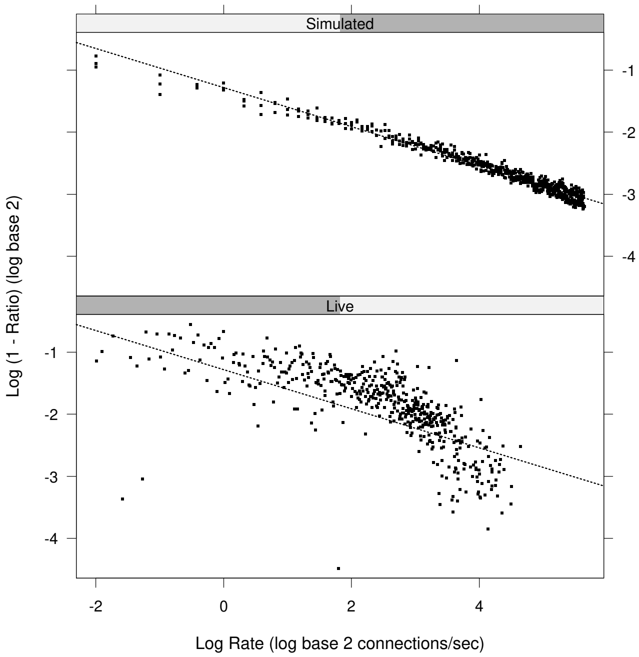

Figure 10: $\log_2(1-\theta)$ is graphed against $\log_2(\rho)$ for the live data in the bottom panel and for the simulated data in the top panel. The line on both panels is the least squares line fitted to the simulated data.

average number of files in the range 10 to 20. However, we did not systematically study this departure from the model since the magnitude is small.

From Equation 9, we estimate σ^2_{bs} by

$$\hat{\sigma}^2_{bs} = \frac{8\Gamma(1/2)}{3\Gamma^2(3/4)}\hat{\sigma}^2_{b\epsilon}.$$

From Equations 10 and 11, we estimate θ_b by

$$\hat{\theta}_b = \frac{\hat{\sigma}^2_{bn}}{\hat{\sigma}^2_{bs} + \hat{\sigma}^2_{bn}}.$$

We study how $\hat{\theta}_b$ depends on $\hat{\rho}_b$ for the 500 blocks and compare this dependence with that predicted by the theoretical study to come. The bottom panel of Figure 10 graphs $\log_2(1-\hat{\theta}_b)$, against $\log_2\hat{\rho}_b$ for the 500 blocks. (The top panel will be described later.) The graph shows that $1-\hat{\theta}_b$ tends to zero as $\hat{\rho}_b$ increases, which means that ℓ_{bj} tends to white noise and the power spectrum approaches a constant. We can see this happening in Figures 8 and 9. In Figure 8, $\hat{\rho}_b$ is low, and the log power spectrum decreases substantially at all frequencies. In Figure 9, $\hat{\rho}_b$ is high, and the log power spectrum is closer to a constant in the sense that for frequencies above 0.1 cycles/inter-arrival, the log power spectrum is nearly constant.

The above methods employ loess smoothing of $z_b(\bar{f}_{bk})$ to get an estimate of the log spectrum, and least squares fitting to $z_b(\bar{f}_{bk})$ to get estimates of the parameters of the model. The term $g(f) = |1-e^{2\pi i f}|$ in Equation 22 introduces substantial curvature in $p_b(f)$ near the origin, so to cope with this curvature, the $z_b(\bar{f}_{bk})$ were smoothed by loess as a function of $\log_{10}(g(\bar{f}_{bk}))$ and then plotted against $\bar{f}_{bk}$; this results in a much better fit than smoothing directly as a function of $\bar{f}_{bk}$. The loess smoothing parameter was 3/4 and the fitting was locally quadratic. The three model parameters were estimated by nonlinear least squares fitting of $z_b(\bar{f}_{bk})$ to $\log_{10}(p_b(\bar{f}_{bk}))$. In both these cases we are invoking ATS: average, transform, and then smooth. In the first case the smoothing is accomplished by loess, and in the second by model fitting. The averaging before taking logs is important; if we proceed without it, as is done in [13], then estimates based on the logs are inefficient, that is, they do not use full information in the data [6]. We fit on the log scale because the large change in the power spectrum due to long-range persistence is smoother than on the original scale, so estimation methods based on the power spectrum perform more reliably.

6.2 Theoretical Study

Just as for the univariate distribution of ℓ_j, connection-rate superposition is used to study the autocorrelation of ℓ_j theoretically. We take as a base point process the model in Section 3.5 with the base rate equal to $\rho_1^* = 0.25$ c/s, a value that is close to the minimum $\hat{\rho}_b$. We generate start times with this rate, and superpose different numbers of them to get processes at higher rates $\rho_k^* = k\rho_1^*$ for $k = 1$ to 100. We then study the results by the same power-spectrum methods used to study the empirical data. In particular, we estimate θ in the same way. In the top panel of Figure 10, the simulation estimates of $\log_2(1-\theta)$ are graphed against the generation values of $\log_2(\rho)$. The line on both panels of the figure is fitted by least squares to the simulation points of the top panel; this fit is the result given in Equation 12. The simulation requires a value of θ for the base process; we estimate this value by choosing the one that results in the closest fit of the simulated values in the top panel to the live data in the bottom panel, just as we formed the estimate λ_1^* in the theoretical study of Section 5. The resulting value of θ is 0.45.

6.3 Discussion

The theoretical and empirical results are in very close qualitative agreement. The estimates of the power spectra have the same behavior and the same change with the rate. For example, the estimates of $1-\theta$ decrease with ρ just as in the live data. Figure 10 shows that for the higher rates, the simulated superposed process begins to depart from the pattern of the data. Since $1-\theta$ is the fraction of the variance due to the persistent series s_j, the departure means that for the simulated data there is more variability due to s_j. But in both cases, the magnitude of the variability is quite small, so the differences are minor.

7. STATISTICAL GENERATION MODEL

The statistical generation model for ℓ_j given in Section 3.5 is developed to reflect the characteristics of the univariate distributions uncovered in Sections 5, and the characteristics of the autocorrelation uncovered in Section 6. The mean and variance of $s_j + n_j$ match the mean and variance of the target extreme-value distribution for ℓ_j, and the autocorrelation function matches the target $a_\ell(k,\rho)$. But the $s_j + n_j$ do not in general have an extreme-value distribution with parameters $\lambda(\rho)$ and $\alpha(\rho)$. For this reason, we transform the $s_j + n_j$ so that the ℓ_j have exactly the extreme-value distribution. The autocorrelation structure of ℓ_j is now not exactly $a_\ell(k,\rho)$; but by generating values of ℓ_j from the model, and estimating their spectra is we did in Section 6, we found the

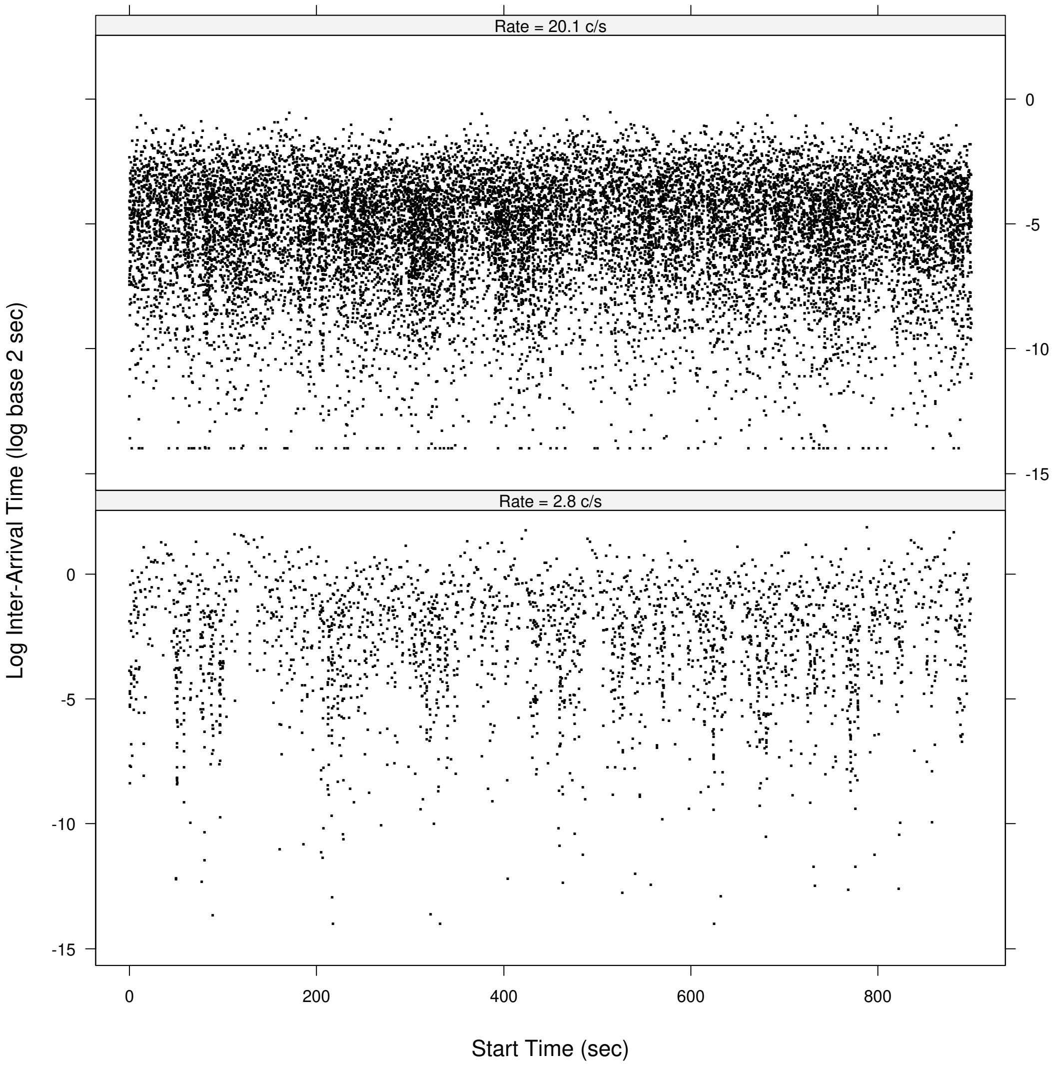

Figure 11: Log inter-arrival time is graphed against start time for two blocks of HTTP connection start times generated from the statistical model. The rates for these synthetic data are the same as the sample rates for the live data in Figure 2. To facilitate comparison of the generated and live data, 82 values of the generated inter-arrivals less than the minimum live time of 2^{-14} sec have been set to the minimum.

autocorrelation is close to $a_\ell(k, \rho)$. Consequently, except for the network artifacts discussed in Section 4, the statistical characteristics of the live HTTP start times and the generated HTTP start times are in close agreement. This is illustrated by comparing Figures 2 and 11. In Figure 11, the panels graph generated start times with values of ρ equal to 2.8 c/s and 20.1 c/s, the same as the sample rates of the live data in Figure 2. We altered the generated times to add a part of the network artifact. The minimum inter-arrival time for the live data is 2^{-14} sec. There are 82 generated inter-arrival times less than the minimum; these values have been changed to the minimum in the figure.

8. DISCUSSION

The results of this paper are given in Section 3. The following are comments on interesting issues that need further work.

The empirical study is based on HTTP requests from one specific network, the Bell Labs network. But because the theory is not specific to the network, we believe that the results and the generation model stand a good chance of holding for other networks. There might be a need for a different calibration of ρ. The calibration might be multiplicative: if ρ^* is the rate for another network, then the statistics for ρ^* are those for the Bell Labs network with rate $\rho = c\rho^*$. A calibration is needed if a connection rate of, say, 20 c/s on another network implies a different number of traffic sources than on the Bell Labs network.

The work here applies to time scales for which there is not an appreciable effect due to daily and weekly variation, 15 minutes or less. However, it would be relatively straightforward to model such variation. In the notation of Section 3, we would model ℓ_{1j} in Equation 1. If the time scale were to be a matter of hours, the model could be a deterministic component that reflected typical change, for example, the change seen in Figure 1. For longer time scales, say days, the model could be a stochastic time series model with periodic components.

The start time generator provides stochastic input to our simulation system for HTTP traffic. However, the stochastic input must encompass more. Accompanying each of the start times must be the size of the request file sent from the client to the server and the size of the downloaded file. The start times are a statistical point process. The times together with the sizes form a statistical marked point process. A next step is to build a model for this marked process, determining the file size distributions, whether there is time correlation in the sizes, and whether the file sizes are correlated with the inter-arrivals.

It is possible to couch the development of the model as a deconvolution problem. We have three random variables ℓ_j, s_j, and n_j. The distribution of ℓ_j has an extreme-value distribution with parameters $\lambda(\rho)$ and $\alpha(\rho)$. The s_j are a Gaussian process with mean 0, variance $\sigma_s^2(\rho)$, and autocorrelation $a_s(k)$. The n_j are i.i.d. with variance $\sigma_n^2(\rho)$. The deconvolution problem is to find a distribution for n_j so that $s_j + n_j$ has a distribution as close as possible to that of ℓ_j and is easy to compute.

9. REFERENCES

[1] P. Barford and M. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *Proc. ACM Sigmetrics '98*, pages 151–160, 1998.

[2] R. A. Becker, W. S. Cleveland, and M. J. Shyu. The Design and Control of Trellis Display. *Journal of Computational and Statistical Graphics*, 5:123–155, 1996.

[3] J. M. Chambers. *Programming with Data*. Springer, New York, 1998.

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

[5] W. S. Cleveland and S. J. Devlin. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83: 596–610, 1988.

[6] W. S. Cleveland, C. L. Mallows, and J. E. McRae. ATS Methods: Nonparametric Regression for Nongaussian Data. *Journal of the American Statistical Association*, 88:821–835, 1993.

[7] D. R. Cox. *Renewal Theory*. Chapman and Hall, 1962.

[8] S. Deng. Empirical model of www document arrivals at access link. In *Proceedings of ICC/SUPERCOMM*, 1996.

[9] A. Erramilli, O. Narayan, and W. Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *IEEE/ACM Transactions on Networking*, 4:209–223, 1996.

[10] A. Feldman, A. Gilbert, W. Willinger, and T. G. Kurtz. The changing nature of network traffic: Scaling phenomena. *Computer Communication Review*, 28, 1998.

[11] A. Feldmann. Characteristics of TCP Connection Arrivals . Technical report, AT&T Labs Research, 1998.

[12] H. Fowler and W. Leland. Local Area Network traffic Characteristics, with Implications for Broadband Network Congestion Management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1145, 1991.

[13] J. Geweke and S. Porter-Hudak. The Estimation and Application of Long Memory Time Series Models. *Journal of Time Series Analysis*, 4:221–238, 1983.

[14] J. R. M. Hosking. Fractional Differencing. *Biometrika*, 68:165–176, 1981.

[15] N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions*. Houghton Mifflin Company, Boston, 1970.

[16] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the Self-Similar Nature of Ethernet Traffic. *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

[17] W. Leland and D. Wilson. High Time-Resolution Measurement and Analysis of LAN Traffic: Implications for LAN Interconnection. In *Proceedings of IEEE Infocom*, 1991.

[18] B. Mah. An Empirical Model of HTTP Network Traffic. In *Proceedings of IEEE Infocom '97*, 1997.

[19] S. McCanne and S. Floyd. NS (Network Simulator). http://www-nrg.ee.lbl.gov/ns/, 1998.

[20] S. McCanne and S. Floyd. UCB/LBNL Network Simulator - ns (version 2). http://www-mash.cs.berkeley.edu/ns/, 1998.

[21] J. Mogul. Network Behavior of a Busy Web Server. Technical Report 95/5, DEC Western Research Lab, 1995.

[22] K. Park, G. Kim, and M. Crovella. On the Relationship Between File Sizes, Transport Protocols, and Self-Similar Network Traffic. In *Proceedings of the IEEE International Conference on Network Protocols*, 1996.

[23] K. Park, G. Kim, and M. Crovella. On the effect of traffic self-similarity on network performance. In *Proc. SPIE Intl. Conf. Perf. and Control of Network Systems*, 1997.

[24] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.