Empirically-Derived Analytic Models of Wide-Area TCP Connections: Extended Report

Vern Paxson
Lawrence Berkeley Laboratory and
EECS Division, University of California, Berkeley
1 Cyclotron Road
Berkeley, CA 94720
vern@ee.lbl.gov
LBL-34086

June 15, 1993

Abstract

We analyze 2.5 million TCP connections that occurred during 14 wide-area traffic traces. The traces were gathered at five "stub" networks and two internetwork gateways, providing a diverse look at wide-area traffic. We derive analytic models describing the random variables associated with *telnet*, *nntp*, *smtp*, and *ftp* connections, and present a methodology for comparing the effectiveness of the analytic models with empirical models such as *tcplib* [DJ91]. Overall we find that the analytic models provide good descriptions, generally modeling the various distributions as well as empirical models and in some cases better.

1 Introduction

Though wide-area networks have been in use since the early 1970's, until recently we have known virtually nothing about the characteristics of the individual connections of different protocols. In the last few years a number of papers have appeared giving statistical summaries of traffic on a per-protocol basis [Càceres89, Heimlich90, CW91, EHS92, WLC92], an important first step. The next step in understanding wide-area traffic is to form models for simulating and predicting traffic.

One such model, *tcplib* [DJ91, DJCME92], is now available. *tcplib* is an *empirical* model of wide-area traffic: it models the distribution of the random variables (e.g., bytes transferred, duration) associated with different protocols by using the distributions actually measured for those protocols at an Internet site.

Ideally we would like to have *analytic* traffic models: simple mathematical descriptions rather than empirical distributions. Such models are easier both to convey and to analyze. The key question is whether analytic models can describe the diverse phenomena found in wide-area traffic as well as empirical models. Our previous work [Paxson91] offered such models but suffered in part from flawed statistical methodology.

In this paper we analyze 14 wide-area traffic traces gathered at seven different sites, five "stub" networks and two internetwork gateways. We derive analytic models describing the random variables associated with *telnet*, *nntp*, *smtp*, and *ftp* connections, and present a methodology for comparing the effectiveness of the analytic models with *tcplib* and with another empirical model constructed from one of the datasets. Table 1 summarizes our main results. Overall we find that the analytic models provide good descriptions, generally modeling the various distributions as well as the empirical models and in some cases better. We develop each of these findings in the remainder of the paper.

In the next section we give an overview of the 14 traffic traces. We describe the gross characteristics of the traces including their traffic mix, and discuss how we filtered the traffic to remove anomalous connections.

The following section presents our statistical methodology. We discuss how we transformed the data and dealt with outliers; our unsuccessful attempts to find "statistically valid" models; the metric we devised for comparing the fit of two different models to a dataset; and our methodology for modeling connection interarrivals, which is more complex than modeling the other random variables associated with a connection.

We then present one section each on modeling *telnet*, *nntp*, *smtp*, and *ftp*. These sections can be read independently if the reader is more interested in one protocol than another, except that the first section describes how to read the plots used in all four sections to compare models. By reading the text accompanying Figures 4, 5, 13, and 30, the reader can if desired skip the remainder of the *telnet* section.

Random variables associated with wide-area network connections can be described as well by analytic models as by empirical models.
When using either type of model, caution must be exercised due to frequent discrepancies in the upper 1% tails.
While in general the analytic models do not match the observed distributions identically in a statistical sense, often a random subsample of hundreds of data points does result in a statistically valid fit, indicating that the analytic models are often *close* though not *exact*.
Bulk-transfer traffic (*ftpdata*, *smtp*, *nntp*, and *telnet* response) is best modeled using log-normal distributions.
Bulk-transfer traffic is not strongly bidirectional; the responses to bulk transfers show little variation relative to the variation in the size of the transfer.
Network traffic varies significantly, both over time and more so from site-to-site, not only in traffic mix but in connection characteristics.
Scaling usually helps significantly in modeling the bytes transferred by *nntp*, *smtp*, *rlogin*, and individual *ftpdata* connections, but is usually not necessary for adequate fits to *telnet* connections and full *ftp* conversations.
Except for *nntp*, connection interarrivals are well modeled using nonhomogeneous Poisson processes with fixed hourly rates.

Table 1: Major Findings

In the last section we summarize the different analytic models and discuss findings in addition to those listed in Table 1. We also include appendices summarizing how we filtered the data prior to analysis, and exploring the effectiveness of modeling *rlogin* traffic using the *telnet* models.

2 Overview of Network Traffic Traces

To develop and then evaluate our models we acquired a number of traces of wide-area traffic. Our main data were from six month-long traces of all wide-area TCP connections between the Lawrence Berkeley Laboratory (LBL) and the rest of the world. With the help of colleagues we also were able to study traces from Bellcore, the University of California at Berkeley, the University of Southern California, Digital's Western Research Laboratory, the United Kingdom–United States academic network link, and traffic between the coNCert[1] network and the rest of the world. We discuss the general characteristics of each of these datasets in turn and then provide summaries of their TCP traffic.

2.1 The LBL Traces

All off-site communication at LBL funnels through a group of gateways that reside on a network separate from the rest of the Laboratory. The first two datasets were taken using a Sun 3/50 residing on the gateway network, using the *tcpdump* packet capture tool [JLM89] running the Berkeley Packet Filter [MJ93]. The Sun 3/50 had kernel modifications to gain a clock resolution of 10 msec. These are the traces discussed in [Paxson91]. When we took the last four traces the monitor workstation had been upgraded to a Sun SLC with a consequent improvement of clock resolution to 1 microsecond.

We used a *tcpdump* filter to capture only those TCP packets with SYN, FIN, or RST flags in their headers, greatly reducing the volume and rate of data (but at the cost of no analysis of intra-connection dynamics). From SYN and FIN packets one can derive the connection's TCP protocol, connection duration, number of bytes transferred in each direction (excluding TCP/IP overhead), participating hosts, and starting time. In principle we could derive the same information using RST packets instead of FIN packets, but we found that often the sequence numbers associated with RST packets were erroneous. Since we could not derive reliable byte counts from RST-terminated connections we excluded them from subsequent analysis.

With this packet capture scheme there are two mechanisms by which packets can be lost. The first is that, if a packet arrives at the Ethernet controller and the controller has run out of kernel memory to buffer the packet, it drops the packet and sets a bit indicating that this event occurred. The Ethernet driver subsequently reads the bit and increments a corresponding counter. It is possible that more than one packet will be dropped before the driver is able to read the bit, so the actual number of dropped packets is unknown but at least as large as the driver's counter.

The second packet-drop mechanism occurs when the kernel determines that the packet filter accepts a packet, but has no

[1] Communications for North Carolina Education, Research and Technology.

more buffer space for saving the packet (due to the user-level program failing to consume previously accepted packets). In this case the kernel drops the packet and increments a counter. Values reported by this counter thus correspond to exactly the number of acceptable packets (in our case, SYN/FIN/RST packets) dropped.

Dataset	Packets (days)	Start	End	Drops
LBL-1	124M (36)	01Nov90	01Dec90	$0 + 0$
LBL-2	?	28Feb91	30Mar91	$0+?$
LBL-3	207M (47)	07Nov91	07Dec91	$9 + 24$
LBL-4	210M (36)	19Mar92	18Apr92	$6 + 233$
LBL-5	337M (35)	24Sep92	23Oct92	$8 + 1808$
LBL-6	447M (31)	24Feb93	26Mar93	$3 + 0$

Table 2: Summary of LBL Datasets

Table 2 summarizes the LBL datasets. The second column gives the total number of network packets received by the kernel for each dataset, along with the number of days spanned by the entire trace. (The statistics missing for the LBL-2 dataset are due to abnormal termination of the tracing program.) Each dataset was then trimmed to span exactly 30 days, beginning at midnight on a Thursday and ending at midnight on a Saturday (i.e., just after 11:59PM Friday night), except for LBL-6, which begins on a Wednesday and ends on a Friday midnight. The "Drops" column gives the drop count reported by the Ethernet driver followed by the drop count reported by *tcpdump*; this last value represents dropped SYN/FIN/RST packets.[2]

Finally, since the LBL datasets span 2.5 years at roughly regular intervals, they provide an opportunity to study how a site's wide-area traffic evolves over time. Such a study is reported in [Paxson93].

2.2 The Additional Traces

As mentioned above, a number of colleagues generously provided access to traffic traces from other sites. The authors of [DJCME92] provided their traces of traffic from Bellcore, U.C. Berkeley, and U.S.C.; Jeffrey Mogul provided traces from DEC-WRL; Wayne Sung provided traces of traffic to/from the coNCert network in North Carolina; and the authors of [WLC92] provided their traces of the UK–US academic network. The first four traces all originate from "stub" sites, while the latter two represent inter-network traffic (though the authors of [WLC92] characterize the UK side of the UK–US traffic as similar to a large stub site since it comprises only a few hosts).

Site	Starting Time	Duration	Drops
Bellcore (BC)	Tue 14:37 10Oct89	13 days	0
UCB (UCB)	Tue 10:30 31Oct89	24 hours	0
USC (USC)	Tue 14:24 22Jan91	26 hours	$\approx 0.6\%^*$
DEC (DEC-1)	Tue 16:46 26Nov91	24 hours	?
DEC (DEC-2)	Wed 17:55 27Nov91	24 hours	?
DEC (DEC-3)	Mon 15:02 02Dec91	24 hours	?
coNCert (NC)	Wed 09:04 04Dec91	24 hours	?
UK-US (UK)	Wed 05:00 21Aug91	17 hours	0

Table 3: Summary of Additional Datasets

The additional datasets are summarized in Table 3. Next to the site name we give in parentheses the abbreviation we will use to identify the dataset. The drop rates for the first three datasets correspond to those listed in [DJCME92]; for the last dataset, to that listed in [WLC92]; and the drop rates for the remaining datasets were unavailable. The USC dataset's drop rate is marked because we found our copy of the trace plagued throughout by "blackouts" of missing packets, occurring almost exactly a minute apart and each blackout lasting roughly ten seconds.[3] Because of these blackouts, we exclude the USC dataset from our interarrival models.

2.3 Filtering of non-WAN traffic

Before proceeding with our analysis we filtered out non-wide-area traffic from the datasets: internal and transit traffic. The details are given in Appendix A. In addition, we removed from the LBL datasets all traffic between LBL and U.C. Berkeley[4]. While traffic with the University forms a significant fraction of LBL's off-site traffic (20-40% of all connections), it is atypical wide-area traffic due to the close administrative ties and the short, high-speed link between the institutions.

2.4 Traffic Overview

We now turn to characterizing the different datasets in order to gauge their large-scale similarities and differences. Of previous traffic studies, only [FJ70], the related [JS69], and [DJCME92] compare traffic from more than one institution. The first two papers found significant differences between their four traffic sites, which they attribute to the fact that the different sites engaged in different applications and had different hardware. The authors of [DJCME92] found that their three sites (which correspond to the USC and UCB datasets in

[2] In the LBL-4 dataset we observed the heated exchange of nearly 400,000 RST packets sent between a lone remote host and three LBL hosts, separated by a geometric mean of 1.3 msec. In LBL-5 we observed the exchange of nearly 120,000 RST packets between a single pair of hosts, virtually all occurring during a 98 second period, separated by a geometric mean of 400 μsec. LBL-6 did not include any RST bursts.

The LBL-5 RST bursts correspond to enough traffic to consume $>$ 500 kbit/sec. We suspect that the RST bursts are the cause of the relatively large number of dropped SYN/FIN/RST packets in LBL-4 and LBL-5.

[3] These blackouts do not correspond to network outages; sequence numbers of TCP connections spanning outages show jumps.

[4] Including *nntp*, unlike [Paxson93], which keeps the *nntp* traffic.

Dataset	# Conn	*nntp*	*smtp*	*ftpdata*	*ftpctrl*	*telnet*	*rlogin*	*finger*	*domain*	*X11*	*shell*	*other*
LBL-1	146,209	40	26	16	3	4	1	4	4	0.2	0.5	0.5
LBL-2	170,718	34	30	16	3	4	1	5	4	0.2	0.2	0.7
LBL-3	229,835	20	33	17	3	4	1	4	11	0.4	0.3	5
LBL-4	449,357	16	21	15	3	2	1	32	5	0.4	0.2	4
LBL-4*	312,429	23	30	21	4	3	1	3	8	0.5	0.3	5
LBL-5	370,397	14	34	22	5	4	1	6	8	0.9	0.2	5
LBL-6	528,784	11	40	23	6	3	0.8	5	5	0.7	0.4	4
BC	17,225	2	49	30	4	4	2	5	0.1	0.1	0.5	2
UCB	37,624	18	45	18	2	2	0.9	12	0.1	0.02	0.2	0.8
USC	13,097	35	27	14	2	3	1	11	2	0.09	0.3	3
DEC-1	72,821	33	35	11	1	0.08	0.05	0.1	20	0	0.001	0.8
DEC-2	49,050	38	22	8	1	0.04	0.06	0.2	29	0	0.02	1
DEC-3	73,440	26	43	9	1	0.07	0.07	0.2	19	0	0.003	1
NC	62,819	1	42	30	4	5	0.3	5	0.8	0.03	0.3	5
UK	25,669	0.02	42	39	7	4	0.4	0.9	1	0.02	0.02	4

Table 4: Percentage Connection Mixes for All Datasets

Dataset	MB	*nntp*	*smtp*	*ftpdata*	*ftpctrl*	*telnet*	*rlogin*	*finger*	*domain*	*X11*	*shell*	*other*
LBL-1	2,852	19	5	65	0.2	6	0.8	0.1	1	3	1	0.1
LBL-2	3,785	14	6	67	0.2	5	1	0.1	0.9	1	3	2
LBL-3	6,710	7	4	67	0.1	4	1	0.1	0.7	3	11	1
LBL-4	11,398	21	4	52	0.1	4	0.9	0.0	0.6	6	10	1
LBL-5	19,269	17	3	57	0.1	3	0.7	0.1	0.4	11	8	1
LBL-6	22,076	22	5	57	0.2	2	0.7	0.1	0.5	8	3	0.8
BC	346	4	8	78	0.3	4	2	0.2	0.1	0.1	2	2
UCB	318	23	16	50	0.3	4	3	0.9	0.0	0.2	0.6	1
USC	362	62	3	18	0.1	2	0.9	0.3	0.3	5	7	2
DEC-1	981	43	17	38	0.2	0.1	0.2	0.0	0.7	0.0	0.0	1
DEC-2	819	54	14	30	0.1	0.0	0.2	0.1	0.6	0.0	0.0	2
DEC-3	1,379	52	16	30	0.1	0.1	0.2	0.1	0.6	0.0	0.0	1
NC	1,553	9	8	68	0.3	5	0.3	0.1	0.3	0.1	0.3	8
UK	625	0.5	11	80	0.4	3	0.5	0.0	0.3	0. 1	0.5	4

Table 5: Percentage Byte Mixes for All Datasets

this paper, as well as part of the BC dataset) had quite different mixes of traffic, but that the characteristics of any particular protocol's traffic were very similar (though they did not quantify the degree of similarity).

Table 4 shows the "connection mix" for each of the datasets. The second column gives the total number of connections recorded, and the remaining columns the percentage of the total due to particular TCP protocols. The mixes for BC, UCB, and USC differ from those given in [DJCME92] because the latter reports *conversation* mixes, where multiple related connections have been combined into single conversations. (The authors also used twenty-minute silences to delimit the end of connections, instead of FIN packets.)

From the Table it is immediately clear that traffic mixes for all protocols vary substantially, both from site-to-site and over time (for LBL). There are also a number of anomalies which merit comment:

- The huge spike in the LBL-4 *finger* connections, the large jump in *other* connections at LBL-3, and the increasing proportion of *ftpctrl* traffic (i.e., the control side of an *ftp* conversation), are all due to the use of background scripts to automate periodic network access. Reference [Paxson93] explores this phenomenon further. LBL-4* shows the LBL-4 connection mix with the periodic *finger* connections removed, as they significantly skew the mix profile.
- The large variance of LBL's *nntp* mix is due to changes in LBL's *nntp* peer servers and differences in the rate at which new news arrives. Again, see [Paxson93] for a

discussion.

- DEC has a "firewall" in place which prohibits traffic other than *nntp*, *smtp*, and *ftp*, and *domain*. The little remaining traffic due to other protocols originated on the outside of the firewall.
- The DEC-2 dataset includes part of the Thanksgiving holiday, accounting for the depressed number of connections.
- As mentioned in [WLC92], the United Kingdom receives its network news from Holland, hence the very low proportion of *nntp* connections.

Table 5 shows the total number of data megabytes transferred (in either direction) for each of the datasets, along with the "byte mix"—the percentage of the total bytes due to each protocol. The LBL datasets show striking growth over time, which we explore further in [Paxson93]. The LBL datasets naturally total more bytes than the others because they span 30-day periods, as opposed to about 1 day for all the other datasets except BC (see Table 3).

We see immediately that, much as with the connection mix, the byte mix also varies considerably both from site-to-site and over time. Some sites (the first three LBL datasets, BC, NC, and UK) are wholly dominated by *ftp* traffic, while others (the last three LBL datasets, UCB, and the DEC datasets) show more of a balance between *nntp* and *ftp* traffic; and USC is dominated by *nntp* traffic. For some sites (UCB, DEC), *smtp* traffic contributes a significant volume, and for others (LBL, USC), traffic due to *X11* and *shell* far outweighs the almost negligible proportion of connections due to those protocols (see Table 4).

We now turn to the development of the statistical methodology that we will use to characterize the individual connections that make up the data shown in Tables 4 and 5.

3 Statistical Methodology

As noted in [Pawlita89], one weakness of many traffic studies to date has been in their use of statistics. Often the studies report only first or perhaps second moments, and distributions are summarized by eye. Frequently they omit discussion of dealing with outliers, and rarely do they report goodness-of-fit methodologies and results. The few cases where goodness-of-fit issues have been discussed are somewhat unsatisfying (the authors of [FJ70] developed their own, apparently never-published goodness-of-fit measure; and in our own previous work [Paxson91] we used the Kolmogorov-Smirnov goodness-of-fit test as a goodness-of-fit *metric*, an inferior choice). We endeavor in this work to address these statistical shortcomings and to present a general statistical methodology that might serve future work as well.

Our initial goal was to develop "statistically valid" analytic models of the characteristics of wide-area network use. By statistically valid we mean models whose distributions for random variables could not be distinguished in a statistical sense from the actual observed distributions of the variables. In this attempt we failed. Most of the models we present do not reflect the underlying data in a statitistically valid sense; that is, we cannot say that our analytic distributions do indeed precisely give the distributions of the random variables they purport to model. We discuss our failure in Section 3.8 below, and then in Section 3.9 develop a "metric" for determining which of two statistically invalid models better fits a given dataset. But first we discuss the value of statistically valid analytic models and our methodology for developing them, as these issues remain fundamental to putting our results in perspective.

3.1 Analytic vs. Empirical Models

For our purposes we define an analytic model of a random variable as a mathematical description of that variable's distribution. Ideally the model has few bound parameters (making it easy to understand) and no free parameters (making it predictive), in which case it fully predicts the distribution of similar random variables derived from datasets other than the ones used to developed the model. But typically the model might include free offset and scale parameters, in which case it predicts the general shape of future distributions but not the exact form. If those parameters are known for a future dataset, then the model becomes fully predictive for that dataset.

In contrast, an empirical model such as *tcplib* describes a random variable's distribution based on the observed distribution of an earlier sample of the variable. The empirical model includes a great number of bound parameters, one per bin used to characterize the variable's distribution function; it may be predictive but not easy to understand.

There are a number of advantages of an analytic model compared to an empirical model for the same random variable:

- analytic models are often mathematically tractable, lending themselves to greater understanding;
- analytic models are very concise and thus easily communicated;
- with an analytic model, different datasets can be easily compared by comparing their fitted values for the model's free parameters.

A key question, though, is whether an analytic model fully captures the essence of the quantity measured by a random variable. An empirical model perfectly models the dataset from which it was derived; the same cannot be said of an analytic model. If the analytic model strays too far from reality, then, while the above advantages remain true, the model no

longer applies to the underlying phenomena of primary interest, and becomes useless (or misleading, if one does not recognize that the model is inaccurate).

The key question then is how to tell that an analytic model accurately reflects reality as represented by a dataset of samples. One approach is to require that the random variable distributions predicted by the model and those actually observed be indiscernable in a statistical sense. To test for such agreement we turn to goodness-of-fit techniques.

3.2 Goodness-of-fit Tests

The random variables we model (amount of data transferred, connection duration, interarrival times, and ratios of these quantities) all come from distributions with essentially unbounded maxima. Furthermore, these distributions are either continuous or, in the case of data transferred, continuous in the non-negative integers. As such the values of the variables do not naturally fall into a finite number of categories, which makes using the well-known chi-squared test less than ideal because it requires somewhat arbitrary choices regarding binning [Knuth81, DS86].

The goodness-of-fit test commonly used with continuous data is the Kolmogorov-Smirnov test. The authors of [DS86], however, recommend the Anderson-Darling (A^2) test [AD54] instead. They state that A^2 is often much more powerful than either Kolmogorov-Smirnov or chi-squared, and that A^2 is particularly good for detecting deviations in the tails of a distribution, often the most important to detect. We followed their recommendation and, in attempting to develop statistically valid models, always used A^2 in assessing goodness-of-fit.

3.3 Logarithmic Transformations

When analyzing data drawn from distributions unbounded in one direction and bounded in the other, often it helps to re-express the data by applying a logarithmic transformation [MT77]. We found that for many of our models logarithmic transformations were required to discern patterns in the large range of values in the data.

For convenience we developed and tested our models using a $\log_2 x$ transformation. Note that, when converting from logarithmic models back to untransformed models, arithmetic means of transformed values become *geometric* means of the untransformed values, and standard deviations become *factors* instead of additive values. For example, a log-normal model with $\bar{x} = 4.0$ and $\sigma = 2.5$ specifies that any observation within a factor of 5.66 ($2^{2.5}$) of 16 ($2^{4.0}$) lies within one standard deviation of the geometric mean. Thus, 2.83 ($= 16/5.66$) and 90.56 ($= 16 \times 5.66$) are the boundaries of values lying within one standard deviation of the geometric mean, which is 16.

3.4 Dealing with Outliers

When applying a logarithmic transformation to non-negative data, one immediately runs into the problem of what to do with data equal to zero. Fortunately for us, in our data such values are rare (and confined to values representing number of data bytes transferred), so we decided to eliminate any connections in which the number of bytes transferred in either direction was zero. We report in Appendix B the number of connections thus eliminated for each dataset; in the worst case they comprised 0.5% of the total connections. An alternative approach would have been to bias our logarithms, by using $\log_2(x+1)$ rather than $\log_2 x$; we rejected this approach as being error-prone when converting to and from the logarithmic models.

Some of our datasets also exhibited values so anomalously large that we removed their associated connections from our study. These outliers were much rarer than those discussed above. Often the values were clearly due to protocol errors (for example, connections in which the sequence numbers indicated $2^{32} - 1$ bytes transferred). We discuss these outliers also in Appendix B.

Finally, we restricted our analysis to datasets with at least 100 connections of interest, to prevent small, anomalous datasets from skewing our results.

3.5 Censored Data

Some of our models describe only a portion of the distribution of a random variable (such as the upper 80% of the distribution). Reference [DS86] discuss modified goodness-of-fit tests (including A^2) to use with such *censored* distributions, in which a known fraction of either tail has been removed from the measurements prior to applying the test. In addition, they describe a method (due to Gupta [Gupta52]) for estimating the mean and variance of such a censored distribution, which can be used to derive estimated parameters of a model from censored data.

3.6 Deriving Model Parameters from Datasets

Often a model has free parameters that must be estimated from a given dataset before testing the model for validity in describing that dataset. For example, a log-normal model may require that the geometric mean and standard deviation be estimated from the dataset. The authors of [DS86] make the important point that estimating free parameters from datasets alters the significance levels corresponding to statistics such as A^2 computed from the fitted model. They then provide both methods to estimate free parameters from datasets, and the required modifications for interpreting the significance of the resulting A^2 (and other) statistics. We followed their approach.

3.7 Model Development vs. Testing

To know if a model is truly predictive, we must test it on data other than that used to develop the model. To this end, we developed all of our models using the first half of the LBL-1 through LBL-4 datasets. We refer to these below as the "test datasets". We then tested the models against the second half of these LBL datasets along with the entirety of the remaining datasets (including all of LBL-5 and LBL-6).

Below we compare our analytic models with two empirical models: one derived from the UCB dataset, which is essentially the same as the *tcplib* model, and one derived from all of LBL-2. Thus, in keeping with our goal of testing models only on data other than that used to develop them, we do not report results for fits to these datasets. An exception is for our interarrival models, which in general we do not compare to the empirical models (see Section 3.11 below).

3.8 Failure to Find Statistically Valid Models

Using the methodology described above, we attempted to develop models for a number of random variables for TCP connections of various protocols. While we often could find fairly simple analytic models that appeared to the eye to closely match the distributions of the random variables for a given dataset, these models rarely proved valid at a significance level of 5%, or even 1%, when tested against other datasets.[5]

What we found tantalizing, though, is that often, when we subsampled the dataset, we *did* find valid fits to the smaller sample. This pattern held whether the subsamples were constructed randomly or chronologically (for example, testing each day in the LBL datasets separately). We tested whether the pattern was due to daily variations in the model's parameters by using autocorrelation plots. We found such patterns only in the arrival process and bytes transferred of *nntp*, and bytes transferred by *smtp* connections. We discuss these findings below in Sections 5.2 and 6.2. We did not find any consistent patterns in the LBL *telnet* or *ftp* test datasets, ruling out simple hourly, daily, or weekly patterns in the parameters.

These findings are consistent with our models being *close* to describing the distributions but not statistically *exact*. In such a case it will take a large number of sample points for a goodness-of-fit test to discern a difference between the distributions. When we subsample we present the test with fewer points and the fit is then more likely to be found valid.

Figure 1 illustrates the problem. Here we see the distribution of $\log_2$ of the bytes sent by the *telnet* responder (i.e., not the host that began the connection) for the first half of the LBL-4 dataset. Fitted against the distribution is our responder-bytes model, which uses a normal distribution for

[5] A significance level of 5% indicates a 5% probability that the A^2 test erroneously declares the analytic model to *not* fit the dataset. A 5% test is more stringent than a 1% test; it errs more often because it demands a closer correspondence between the model and the dataset before declaring a "good fit." See [Ross87, pp. 205-206] for further discussion of significance levels.