# Growth Trends in Wide-Area TCP Connections*

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

Revised May 11, 1994

## Abstract

We analyze the growth of a large research laboratory's wide-area TCP connections over a period of three years. Our data consisted of eight month-long traces of all TCP connections made between the site and the rest of the world. We find that many TCP protocols exhibited exponential growth in the number of connections made and bytes transferred, even though the number of hosts at the site only grew linearly. While the exponential growth of some of the protocols began tapering off with the final datasets, relatively new information-retrieval protocols such as Gopher and World-Wide Web exhibited explosive growth during the same time. Our study also found that individual users greatly affected the site's traffic profile by the inadvertent or casual initiation of multiple, periodic wide-area connections; that exponential growth is fed in part by more users "discovering" the Internet and in part by existing users increasingly incorporating use of the Internet into their work patterns; and that wide-area traffic geography is diverse and dynamic.

## 1 Introduction

To properly design future networks, we need a thorough understanding of how network use changes and grows with time. Previous studies [Kleinrock76, Quarterman90, Lottor92, Adams93, Merit94] have found that many aspects of network use grow exponentially with time, at least until reaching the carrying capacity of the network. These studies all summarize network "backbone" use of some sort. Because the number of host computers connected to these backbones also grows exponentially, we cannot readily extrapolate the growth of network use by individual sites from the backbone growth: no site-growth, linear site-growth, or exponential site-growth are all consistent with exponential backbone growth.

To our knowledge, no studies have appeared tracking the evolution of a site's wide-area network use over time.[1] In this paper we analyze eight one-month traces of a single site's wide-area TCP connections, spanning altogether three years and more than three million connections. The key question for such a study is: Does traffic at individual sites also grow exponentially, and if so, what factors contribute to the growth?

Such questions cannot be answered by studying a single site, as there is no foundation for assuming that such a site is representative of Internet sites as a whole. But a single site study provides a beginning for exploring the questions of site growth more fully. With this limitation in mind, Table 1 summarizes our major findings, which we develop in the body of the paper. The first finding states that the site traffic did indeed grow exponentially, exceeding the rate at which the site added computing resources, but the second finding tempers this result by suggesting that in some areas the growth will begin tapering off. The third finding indicates that new protocols may step in to take up the slack as the growth of older protocols diminishes. The remaining findings address the roots of the growth: part is due to increasing inadvertent or casual use of the network; part is due to an increasing number of existing computers using the Internet; part is due to increasing use of the Internet by individual users; and part is due to taking advantage of the widening Internet connectivity.

In the next section we review existing statistics and studies of network growth, which show that network traffic generally grows exponentially with time, at least until the network carrying capacity is reached. We then describe how we captured and reduced the data used in our study. The following sections address the points made in Table 1: the overall growth in the site's wide-area traffic; the appearance of periodic traffic; the growth in network use by individual computers or users; and the changing geographic profile of the traffic.

The final section summarizes the implications and limita-

*Appeared in *IEEE Network*, 8(4), pp. 8-17, July/August, 1994.

[1] The closest available statistics are those published by Merit, Inc. [Merit94], reporting aggregate bytes transferred into and out of each NSFnet stub network.

| |
|---|
| At a site where the number of Internet hosts increased 30%/year, wide-area TCP traffic for a number of protocols grew significantly faster, both in the number of connections made and (at even higher rates) the amount of data transferred. For example, email (*smtp*) connections grew 70%/year, and *ftp* data bytes grew 100%/year. |
| The most recent datasets indicated decreasing growth in connection and transfer rates for high-bandwidth protocols such as *shell* and *X11*, consistent with exponential growth becoming *logistic* due to reaching carrying capacity limitations. |
| Use of information-retrieval protocols such as World-Wide Web and Gopher grew extremely rapidly. Over a period of two years, World-Wide Web traffic grew by a factor of *300* per year, and it is now one of the site's dominant protocols. In general, new protocols can exhibit explosive growth when first introduced, significantly affecting a site's traffic profile. |
| The site's traffic profile was significantly affected by periodic Internet connections created either unknowingly or quite casually through the use of background scripts. |
| The site's exponential growth in TCP connections was fed in part by new users "discovering" the Internet, and in part by existing users increasingly incorporating use of the Internet into their work patterns. |
| The geographic profile of wide-area traffic is diverse and dynamic. While at any given time particular states and countries dominate the traffic's geography, this profile changes greatly over time. |

Table 1: Major Findings

tions of our results.

## 2 Prior Work

Many existing collections of network growth statistics show that network traffic grows exponentially with time:[2]

- Kleinrock [Kleinrock76] includes a discussion of ARPAnet growth from October, 1971, through March, 1975. For the first half of this period, traffic grew exponentially, rising from $10^5$ packets/day in October, 1971, to 30 times that volume in August, 1973, less than two years later. After August, 1973, though, traffic growth leveled off to about 25% per year. That growth leveled off at this point is not surprising, since the network was then operating at close to its carrying capacity.

- Lottor [Lottor92] reports on how the number of Internet hosts grew from August, 1981, up till January, 1992. While the Internet grew exponentially over this period, the rate lessened with time. Fitting an exponential to the entire body of data yields a growth rate of 140%/year. During the last three years of the study, growth diminished to 100%/year (from 80,000 hosts at the beginning of 1989 to 727,000 at the beginning of 1992).

  More recent data [Lottor93] shows this growth rate leveling off further still, with 2,056,000 hosts at October, 1993. The last three years of growth correspond to 85%/year.

- Statistics available from Merit, Inc. [Merit94], show that NSFnet backbone traffic has been growing exponentially, from $1.3 \cdot 10^{12}$ bytes/month in March, 1991, to $1.1 \cdot 10^{13}$ bytes/month in March, 1994. This growth corresponds to an increase of about 105%/year. During this same time, the number of networks connected to the NSFnet has risen exponentially from 2,501 to 28,578, an overall growth rate of about 120%/year, though over the most recent eight months the rate has climbed to about 190%/year. Unfortunately, the number of different *hosts* making connections over the backbone is not available.

  The NSFnet T1 backbone traffic is further studied by Claffy et. al. [CPB93], who found that the number of bytes traversing the T1 backbone grew quadratically between June, 1988, and June, 1992, though this growth trend is at least partially influenced by the traffic switching over to the T3 backbone later in the study period.

- Statistics for USENET network news traffic [Adams93, Quarterman90] show steady exponential growth since October, 1984. Traffic volume grew from $4 \cdot 10^5$ bytes/day to $5.8 \cdot 10^7$ bytes/day (as of September, 1993). These totals exclude news article headers, which add a fairly constant 25% overhead to the volume. Over this nine-year period, traffic grew at a steady rate of nearly 80% a year, as shown in Figure 1. This constant growth rate sustained over nearly a decade is striking.

[2] Plots of many of the statistics discussed in this section can be found in [Quarterman93].

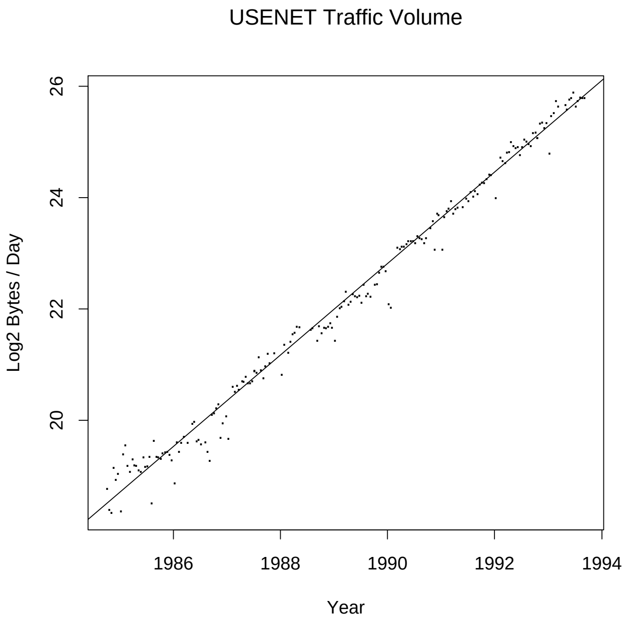

Figure 1: Exponential Growth in USENET Traffic Volume

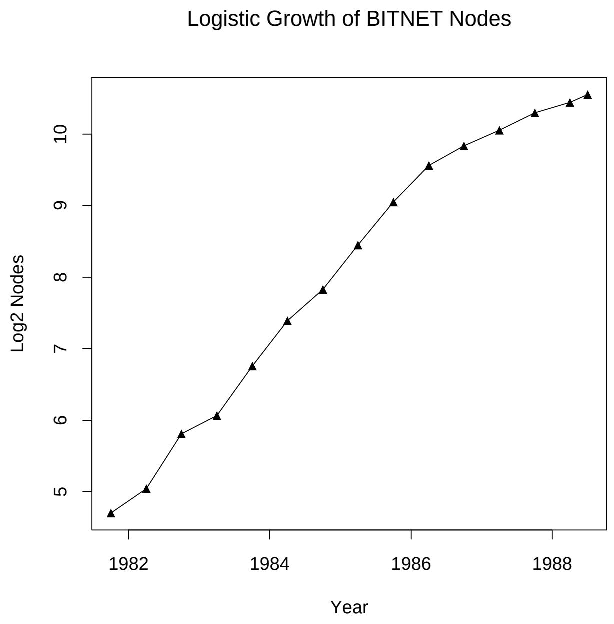

Figure 2: Logistic Growth in BITNET Nodes

During the same interval, the number of sites carrying USENET news also grew exponentially, from 520 to 31,747, but at a slower rate of about 60%/year, consistent with an increase of about 10%/year in the news traffic generated by each site.

- Monthly statistics for the number of hosts in the European RIPE network [Terpstra93] show exponential growth of about 100%/year for the period from January, 1992, through October, 1993. Prior to that time the growth is considerably more rapid, about 260%/year, but also much more fitful.

Exponential traffic growth must at some point slow down. The Kleinrock study shows such a turndown clearly, as does the declining rate of the Internet host count (but not, at least yet, the USENET or NSFnet statistics). Some researchers propose modeling network traffic growth with a *logistic* distribution[3] rather than an exponential, to take into account the carrying capacity of the network [Solensky92] and saturation of demand [Gurbaxani90]. In this regard, we should be alert to the presence of inflection points in growth curves, as these may correspond to approaching the current carrying capacity or demand limits.

Figure 2 shows the number of BITNET nodes over a seven year period, taken from [Gurbaxani90]. The X-axis gives the year and the Y-axis $\log_2$ of the number of BITNET nodes. The inflection point at 1986 shows the characteristic slow-down of logistic growth as network demand or capacity begins to saturate.

## 3 The Site Data

For our study we collected eight month-long traces of all wide area TCP connections between the Lawrence Berkeley Laboratory (LBL) and the rest of the world, taken between November 1990 and April 1994. The acquisition of the traces, which we summarize here, is discussed in greater detail in [Paxson93].

The University of California operates LBL, which is dedicated to basic research, under contract with the U.S. Department of Energy. During the time period covered by this study, LBL as an institute experienced little growth: staffing levels grew from 2,558 full-time equivalents to 2,681, an increase of about 2%/year, and funding rose from $256 million/year up to $290 million and back down to $262 million.

The traces were captured using the *tcpdump* packet capture tool [JLM89] running the Berkeley Packet Filter [MJ93]. 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. From SYN and FIN packets one can derive the connection's TCP protocol, connection duration, number of bytes transferred in each direction, participating hosts, and starting time.[4]

[3] The logistic distribution is defined by

$$F(x;\alpha,\beta) = 1/(1+\exp(-(x-\alpha)/\beta)).$$

[4] 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.

| Dataset | Pkts. (days) | Start | Finish | Drops |
|---|---|---|---|---|
| LBL-1 | 124M (36) | Thu 01Nov90 | Sat 01Dec90 | 0 + 0 |
| LBL-2 | ? | Thu 28Feb91 | Sat 30Mar91 | 0 + ? |
| LBL-3 | 207M (47) | Thu 07Nov91 | Sat 07Dec91 | 9 + 24 |
| LBL-4 | 210M (36) | Thu 19Mar92 | Sat 18Apr92 | 6 + 233 |
| LBL-5 | 337M (35) | Thu 24Sep92 | Sat 23Oct92 | 8 + 1808 |
| LBL-6 | 447M (31) | Wed 24Feb93 | Fri 26Mar93 | 3 + 0 |
| LBL-7 | 560M (32) | Thu 16Sep93 | Sat 15Oct93 | 0 + 7959 |
| LBL-8 | 735M (30) | Thu 30Mar94 | Sat 30Apr94 | 1 + 482 |

Table 2: Summary of Datasets

Table 2 summarizes the 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.[5] Each dataset was then trimmed to span exactly 30 days. 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. As noted in [Paxson93], the increasing number of dropped packets in the later datasets appears correlated with "RST storms"—periods during which two hosts furiously exchange RST packets.

We reduced the traces by extracting only full TCP connections, i.e., two exchanged SYN packets followed by two exchanged FIN packets. We discarded connections that failed to include both pairs of SYN and FIN packets.

Because of the close administrative ties and the short, high-speed link between LBL and the University of California at Berkeley (UCB), traffic between the two institutes is likely to be atypical wide-area traffic, so we also removed these connections (comprising 20-40% of all connections). We made one exception, keeping LBL-UCB *nntp* traffic; by including all of LBL's *nntp* peers, we can study the net inflow and outflow of network news, the total rates of which should be unaffected by the close ties between LBL and UCB.

We also removed connections that transferred no data, and those that purported to transfer implausibly large amounts of data[6], attributing the latter to protocol errors. Details of the removed connections are given in [Paxson93]; the number removed was almost always less than 1% of a protocol's connections.

Finally, to simplify the analysis and presentation of the data, we aggregated into an "other" protocol those connections that did not account for at least 500 connections during two different months.

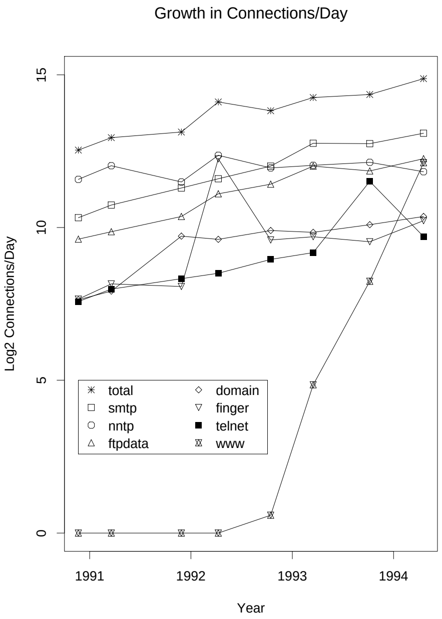

Figure 3: Daily Connection Rate for Most Popular Protocols

## 4 Growth in Traffic Volume

In this section we discuss the overall growth in traffic volume for the different TCP protocols. Figure 3 shows the average daily connection rate for the seven most popular protocols, along with the total connection rate for all of the protocols. The X-axis gives the year, and the Y-axis plots $\log_2$ of the average number of connections made each day.

The varying spacing between the lines of the different protocols immediately shows that the traffic "mix" varies considerably with time. This observation complements that made in [DJCME92] and [Paxson93] that traffic mix also varies significantly from site to site.

Addressing now each protocol in turn:

- The number of *smtp* (electronic mail) connections grew at an approximately exponential rate. A line fitted in a least-squares sense to all of the datasets except LBL-1 and LBL-6 (which are not collinear with the others) gives growth of 70%/year.

- *ftpdata* traffic (corresponding to the data-transfer and directory-listing portion of an *ftp* file-transfer session) is quite noisy, in part due to the presence of weather-map scripts (discussed in Section 5 below). If we remove weather-map traffic, a least-squares fit to all but the fourth dataset gives exponential growth of about 70%/year.

- The considerable variation in the *nntp* (network news) connection rate is due to at least four factors other than just growth in USENET traffic. First, LBL has two *nntp*

[5] The statistics missing for the LBL-2 dataset are due to abnormal termination of the tracing program; this termination, however, did not imply any extra-ordinary loss of packets during the 30-day study period.

[6] Typically close to $2^{32}$ bytes; after removing these connections, the largest connection remaining in any of the datasets was 447 MB.

servers, one primary and the other secondary. These servers sometimes independently connect with their outside peers to receive news, and sometimes receive their news from the other LBL server; the former results in two WAN connections per incoming batch of news, the latter only one. Hence the intra-LBL news-propagation dynamics considerably influences LBL's external *nntp* connection rate. For example, the proportion of the *nntp* connections involving LBL's primary server varied from 58% (LBL-3) to 86% (LBL-4). During LBL-8, the secondary actually took part in 52% of the connections and the primary only 48%.

The second factor is that the rate at which new news arrives depends heavily on the configuration of the remote *nntp* peers. For example, the proportion of connections between LBL and its UCB peers varied from 37% (LBL-1) to 57% (LBL-5).

The third factor is that LBL *nntp* servers keep their connection to their peers open for one minute after they last sent data to the peer, in the hopes that new news will arrive in the interim and can be propagated without requiring a new connection set-up. Thus if new news tends to arrive within a minute of any earlier news, the total number of connections will be lower (other things being equal). One measure of this tendency is the proportion of *nntp* connections that were *inbound* (originated by a remote peer), which will be high if LBL's servers tend to coalesce multiple outbound news batches into a single connection, and low if when propagating news LBL's servers tend to use multiple connections. This rate varied from a low of 15% (LBL-3) to a high of 45% (LBL-2).

The final factor is the proportion of "failed" *nntp* connections. As explained in [Paxson93], an *nntp* connection during which the originator transfers exactly 6 bytes corresponds to a "failure" in the sense that the responder stated it was unable to accept news at that time. Such failures are likely to lead to repeated connections as the originator later tries to again propagate the news to the responder. The failure rate fell steadily from 38% (LBL-1) to 2-6% (LBL-4 through LBL-8).

- *domain* (Internet domain name service) traffic increased greatly between LBL-2 and LBL-3, and otherwise shows fairly flat growth. The large increase is due to the addition of a probably-misconfigured remote peer, as the LBL-3 through LBL-8 traffic is all dominated by connections to just one remote site.

- *finger* (remote user lookup) traffic shows a great deal of variation, including a huge spike in the LBL-4 dataset that returned to an elevated level in LBL-5. The causes for these increases are discussed in Section 5 below.

  We also investigated whether the higher level of *finger* traffic might be due to use of resource discovery tools such as *Netfind* [ST91]. For each dataset we checked the *finger* connections of the ten most popular remote (non-LBL) hosts to see with how many different LBL hosts they connected. We deemed a remote host connecting to more than 25 LBL hosts as engaging in resource discovery. We found no such instances in the first three datasets, one in LBL-4, two in LBL-5 and LBL-6, four in LBL-7, and none in LBL-8. Of the nine resource discovery instances, six involved hosts from the same university as the authors of *Netfind*, indicating our heuristic did successfully identify use of resource discovery tools. Assuring minimal network load was one of the goals of the authors of [ST91]. We found that during the busiest resource-discovery dataset, LBL-7, resource discovery accounted for 8% of all *finger* connections and 3% of all *finger* bytes, providing evidence that this goal has been met.

- *telnet* (remote terminal login) growth grew at about 53%/year, except for a huge increase with LBL-7, all due to a single pair of hosts (again, see Section 5 below).

- *www* (World-Wide Web HTTP) [BlCGP92] traffic absolutely exploded during the final five datasets, sustaining growth of 300-fold/year over a two year period, with no immediate signs of slowing down. For further discussion, see the analysis of *gopher* traffic below.

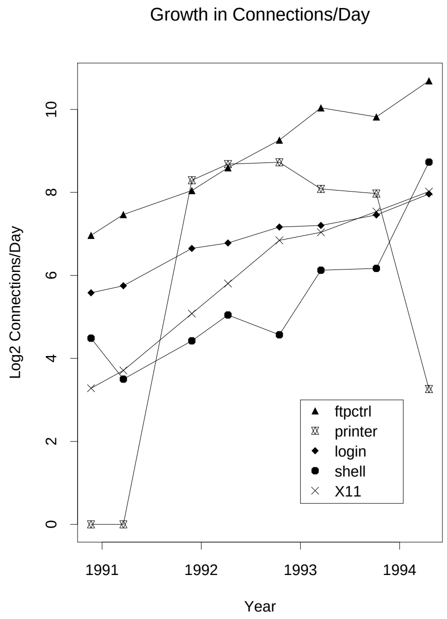

Figure 4: Daily Connection Rate for Less Popular Major Protocols

Figure 4 shows the connection rates for the remaining protocols that had connections during all of the datasets. The Y-axis is again logarithmic:

- *ftpctrl* (the control side of an *ftp* session) shows strong exponential growth of over 100%/year, though the 6th and 7th datasets vary considerably from this trend. It turns out that for the 4th through 7th datasets, *ftpctrl* connections are dominated by weather-map connections (as discussed in Section 5 below). If we remove these connections, the first seven datasets fit well to growth of 66%/year, but the LBL-8 dataset contains almost double the number of connections as predicted by this trend; for a possible explanation of this discrepancy, see the discussion of *gopher* traffic below.
- *printer* (remote printer access) connections increased enormously between LBL-2 and LBL-3, and then leveled off and began heading down again, a phenomenon again discussed in Section 5 below.
- *login* (Unix remote login) traffic grew fitfully, much as did *telnet*, but roughly corresponds to growth of 60%/year.
- *shell* (Unix remote command execution) traffic is very sporadic (but see the discussion of bytes transferred, below). The increase by a factor of 6 between LBL-7 and LBL-8 is entirely due to a single pair of hosts (Section 5). Without this pair of hosts, the number of connections would have *dropped* 35% between the two datasets.
- *X11* (X11 network window system) traffic, on the other hand, shows very consistent exponential growth between LBL-2 and LBL-5, growing at the ferocious rate of 300%/year. With the final four datasets, however, the growth drops to about 75%/year, suggesting that *X11* traffic is exhibiting logistic growth (see the discussion of *X11* traffic volume in bytes below).

Figure 5 shows the connection rates for "new" TCP protocols: those that had no connections during LBL-1[7], other than *www*, which is shown in Figure 3. We also plot as an aggregate "other" protocol those connections of various protocols that did not account for at least 500 connections during two different months.

*X500* refers to the X.500 Directory Services protocol [WRH92]; the protocol exhibited very rapid growth, about 165%/year, but declined between LBL-7 and LBL-8, suggesting that its use may have peaked.

The *gopher* document search and retrieval protocol [AML+93] showed even more dynamic growth: the final five datasets fit fairly well to growing by nearly a factor of *40* each year. But, as shown above in Figure 3, another information retrieval protocol, *www* (World-Wide Web), grew even faster, around a factor of *300* a year! Note that it is very likely that the recent growth in *www*, *gopher*, and *ftpctrl* connections (Figure 4) is related, because *www* connections often lead in turn to *gopher* and *ftpctrl* connections as users follow hypertext links to retrieve new documents.

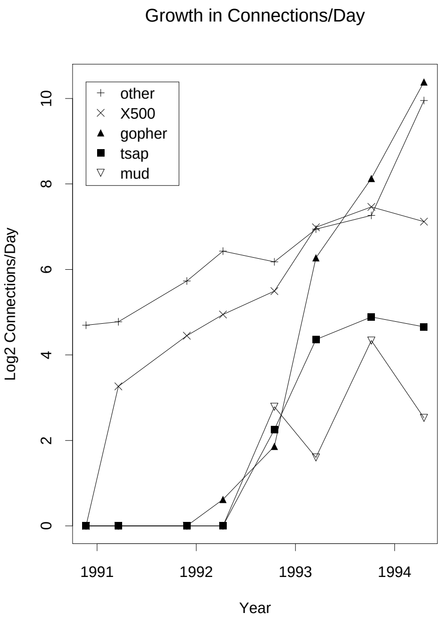

Figure 5: Daily Connection Rate for New Protocols

Both the *www* and *gopher* trends are based on only five datasets and surely must taper off soon, but they dramatically demonstrate how explosively a new type of traffic can grow over a short period of time. See [Rutkowski93] for a related discussion of how the use of new protocols has been growing on the NSFnet backbone, and [SEKN92] for a look at the workings of resource discovery tools such as *X500*, *gopher*, and *www*.

*tsap* refers to the ISO TSAP protocol for layering ISO networking applications on top of TCP [CR86]. The final turndown between LBL-7 and LBL-8 suggests that, along with X.500, this ISO protocol's use is waning.

*mud* refers to a multi-user network game. Its growth is fitful. We simply note that such games are not a new phenomenon: a study of network traffic across the UK-US Academic Network link in August, 1991, attributed 11% of all packets to games [WLC92].

Finally, as mentioned above, *other* traffic aggregates all the remaining connections whose protocols we did not individually analyze. In LBL-7, for example, we observed about 2,750 distinct TCP responder ports in the *other* connections. We only attempted to identify the most popular. In comparison, [Rutkowski93] reports 1,066 different identifiable services present on the NSFnet backbone during May, 1993. We also note that the rapid growth of such connections ($\geq$ 90%/year) is in keeping with the finding in [CPB93] that the "other" protocol traffic on the NSFnet T1 backbone steadily increased.

We now turn to the number of data bytes (sent in both direc-

[7] For this plot only, the protocol counts for LBL-3 were linearly extrapolated from the first 21 days of the dataset.

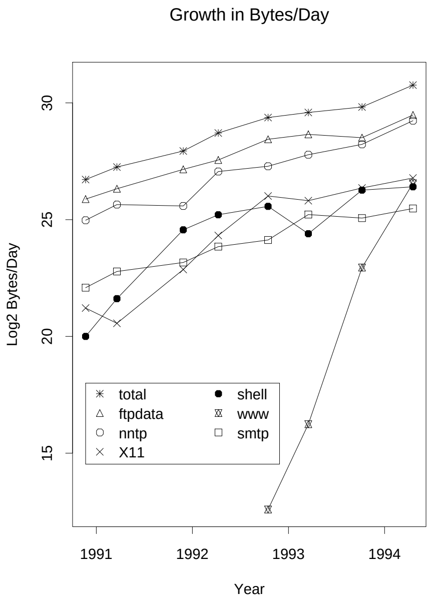

Figure 6: Bytes/Day for Largest Protocols

tions, and excluding TCP/IP headers) per day due to different protocols. Figure 6 plots $\log_2$ of the bytes per day on the Y-axis versus the year on the X-axis:

- The overall growth rate has remained fairly close to 120%/year; that is, *traffic more than doubles each year*.
- Until the last dataset, traffic volume is dominated by *ftpdata*, which during the first six datasets grew exponentially at a rate of about 135%/year. If we remove traffic associated with weather-maps (see below), the datasets still show growth of roughly 100%/year. In comparison, between November, 1992, and May, 1993, the NSFnet *ftpdata* traffic volume grew at a rate of 109%/year [Rutkowski93].
- *nntp*, consistently the second greatest contributor to wide-area bytes, grew fitfully. That the traffic does not reflect the clear exponential USENET growth discussed in Section 2 above is no doubt in part due to the varying "mix" between how often LBL's two *nntp* servers received fresh news separately versus from one another (see the discussion of the *nntp* connection rate earlier in this section). Another major variable in the *nntp* byte rate is the success with which LBL's *nntp* servers propagate news to their remote peers. If mostly successful (i.e., LBL peers tend to have "fresh" news), then the bulk of the USENET traffic will travel both into LBL and then out again, multiplying the total byte count by the number of remote peers to which LBL servers promulgate the articles.

  The overall fit to the *nntp* traffic gives a growth rate of about 130%/year, substantially higher than the USENET growth of 80%/year, and about equal to the annual NSFnet growth rate given in [Rutkowski93].
- The high proportion of data bytes due to *shell* connections is quite startling in light of the low number of daily *shell* connections (see Figure 4 above). We find that *shell* connections vary enormously. For example, in LBL-5 26% of the *shell* connections transferred fewer than 100 bytes, while 2.6% transferred more than 10 MB.

  While the initial growth of *shell* data bytes between LBL-1 and LBL-3 corresponds to growth of 2100%/year, the growth appears to have approached some sort of ceiling, consistent with logistic growth, perhaps due to network bandwidth limits (see the discussion of *X11* traffic below).
- We fit the *smtp* bytes per day to exponential growth of about 100%/year. In comparison, the recent NSFnet growth rate has been 126%/year [Rutkowski93].
- *X11* traffic showed impressive exponential growth between LBL-2 and LBL-5, increasing by more than a factor of 10 each year, much higher than the connection growth of a factor of 4 each year. This difference is due both to an increase in the average connection size (in LBL-2, the geometric mean connection size was 11KB; in LBL-5, 26KB) and to an increasingly heavy upper tail (for example, in LBL-2 only 1% of the connections transferred more than 1 MB, while in LBL-5, 20% did).

  The dip in *X11* traffic volume for LBL-6 and subsequent tepid recovery, however, suggests that *X11* traffic hit an upper-bound and is now responding to logistic pressures. The pattern here is quite similar to that for *shell* traffic, and we again speculate it is due to reaching network bandwidth limitations.
- The volume of *www* traffic grew extremely rapidly—by about *a factor of 750 per year* over the last four datasets—even faster than the number of *www* connections (Figure 3). At this pace, *www* traffic will surpass the volume of *ftpdata* traffic in four more months! The possible inflection point at LBL-7, however, may indicate that the rate is decreasing. Still, it seems likely that within a year *www* traffic will be comparable in total bytes transferred to *ftpdata* traffic.

  We should keep in mind, though, that had we studied only the first five datasets, we would have made similar predictions for *X11* traffic. Its volume leveled off before reaching levels comparable to that of *ftpdata*.

# 5 Anomalous Periodic Traffic

In the later datasets we find large numbers of TCP connections occurring at periodic intervals between the same two hosts, for TCP protocols that do not naturally include periodic communication[8]:

- The huge number of LBL-4 *finger* connections shown in Figure 3 is due to a single user who ran background scripts to query a remote site to see whether a colleague was logged in there. These scripts resulted in 136,928 connections, averaging one every 20 seconds. The *finger* connections dropped considerably in LBL-5, to a total of 23,122 connections, but of these, 6,329 occurred between a single (different) pair of hosts during a five hour period, each connection arriving about a quarter second after the previous one completed.

  In LBL-6 a similar pattern again appears, with 6,256 connections between a single (still different) pair of hosts, almost all over a two-day period. At first the connections arrived about 22 seconds apart. Then evidently a second script was started, as the connection interarrivals varied between 0 and 22 seconds but adjacent interarrivals summed to 22 seconds. Ultimately a third and then a fourth script ran, each drifting in and out of phase with the others.

  Of the 35,868 *finger* connections in LBL-8, 33% were due to a single LBL host periodically querying seven remote hosts, typically about once every 90 seconds. Thus, one LBL host accounted for virtually all of the growth between LBL-7 (which did not have any particularly busy host or pair of hosts) and LBL-8.

- The huge jump in *printer* connections between LBL-2 and LBL-3 (Figure 4) is due to the use of background scripts to query a remote printer queue.

- The exponential growth in *ftpctrl* connections (Figure 4) is fed mainly by the use of background scripts to periodically fetch weather maps (satellite images) from sites in Colorado and Illinois[9]. Use of these scripts began during LBL-3 (932 connections) and grew rapidly during LBL-4 (3,723 connections), LBL-5 (8,533 connections), and LBL-6 (19,264 connections, 61% of the total). The rate fell to 10,746 connections during LBL-7, due at least in part to our raising user awareness of the large impact the scripts had made during LBL-6.

- In LBL-6 we observed 1,988 *telnet* connections (11% of the total) between the same two hosts, all transmitting 3 bytes from the originator to the responder and 175 bytes in the other direction, and lasting about 175 seconds. An amazing 35% of these connections arrived between 180.434 and 180.438 seconds apart.

  In LBL-7 we observed 69,547 connections (79% of the total) between another (different) pair of hosts. Both the LBL-6 and LBL-7 spikes turned out to be due to a *telnet* dial-up server that when contacted automatically attempted to connect with a remote library catalog service. The server suffered from a hardware problem that made it continually believe someone had dialed up, so it perpetually generated a new *telnet* connection as soon as the previous attempt timed out. This problem appears to have remained unnoticed for many months.

- 89% of the LBL-8 *shell* connections occurred between a single pair of hosts. Connections came heavily clustered on the hour and half-hour. These connections turned out to be generated by two scripts, one for mirroring a database between LBL and the remote site, and one for batching up network news in lieu of a proper news feed.

The *ftp*, *finger*, and *printer* connections were apparently initiated quite casually by the users involved. For example, in LBL-4 we observed a single host fetching weather map scripts to have four-to-five scripts running simultaneously. These managed to synchronize with one another and we observed multiple connections virtually identical in bytes transferred and duration, repeating every half hour for days on end.

The main lessons we draw from observing this traffic are that (1) sites would greatly benefit from routine monitoring of their traffic patterns. If Internet services incurred charges on a per-connection basis, some of these spikes would have proven extremely expensive; and (2) networks must be engineered to be resilient in the face of periodic traffic, which can otherwise lead to global synchronization [FJ93].

# 6 Per Capita Network Use

In this section we look at how wide-area network use has grown on a per capita (i.e., per computer or per user) basis, in attempt to discern the different factors contributing to site-wide exponential growth. We first note that while the number of Internet hosts has been doubling every year, LBL's hosts have been growing either linearly, or at a relatively modest exponential rate. Figure 7 shows the increase in the number of unique Internet addresses registered in LBL's database over the time spanned by the datasets. The solid line corresponds to growth of 30%/year, and the dotted line to a linear increase of 765 hosts/year. Recall from the beginning of Section 3 that LBL's personnel and budget have been relatively flat over this same time period; thus the added hosts correspond to the increasing computerization of a fixed population.

Figure 8 shows the number of local hosts that took part in at least one wide-area connection during each 30-day period. (Here *ftp* refers to *ftpctrl* connections.) We call such a host an *active* host. The dashed lines show exponential fits for *telnet* and *ftp*, and a linear fit for *smtp*. The *telnet* fit corresponds to active-host growth of 44%/year, though it is fairly rough over the final four datasets. The *ftp* exponential fit of

[8] Unlike *domain*, for example.

[9] And later Missouri, when the Illinois site stopped offering this service.

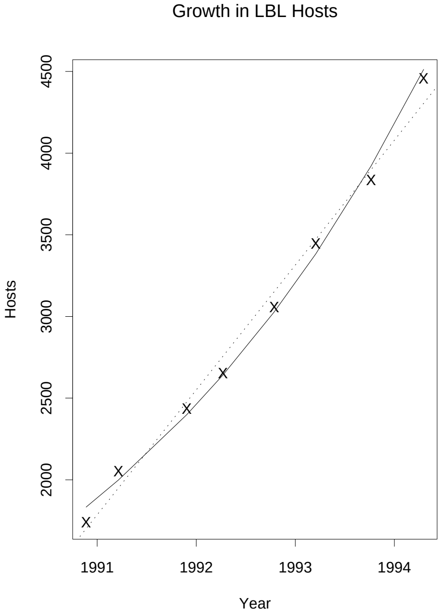

Figure 7: Growth in LBL Hosts

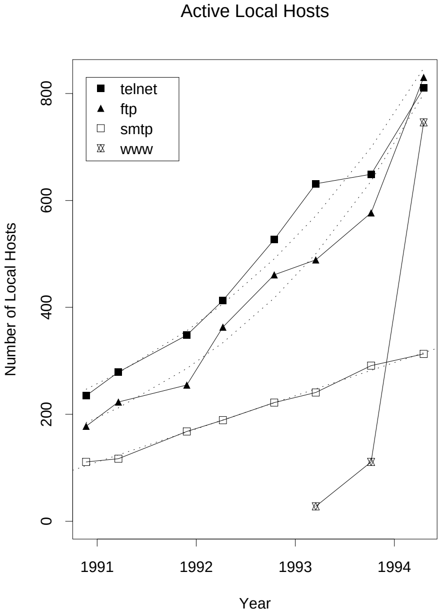

Figure 8: Participation in Wide-Area Connections

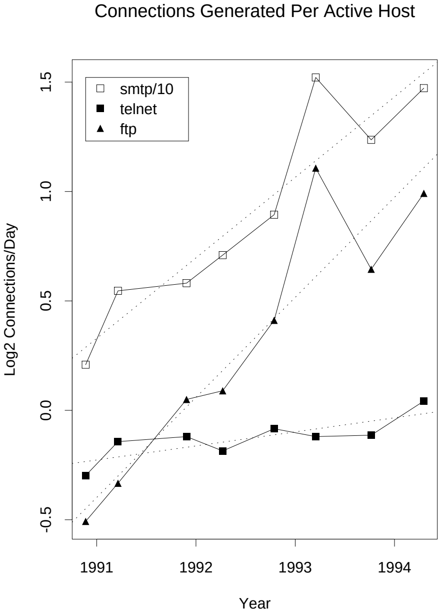

Figure 9: Daily Connection Rate of Active Hosts

54%/year is even rougher. The *smtp* linear fit of 62 active hosts/year, on the other hand, is quite persuasive. Finally, the *www* growth dramatizes how use of World-Wide Web has spread very rapidly through LBL during the last year of our study.

We now look at the growth in connections generated by individual hosts. We first note that our datasets support the "busy-source" and "favorite-site" effects, first noted by [Kleinrock76] and later confirmed by [DJCME92, CPB93], that only a handful of hosts dominate network traffic. For *telnet*, the 5 busiest local hosts generated between 25-40% of all connections; for *ftp*, about 40-50%; and for *smtp*, about 50-70%.

We next turn to gauging whether individual users are increasingly using wide area connections. We do this by looking at "per capita" connection rates. We computed the average number of daily connections made by each active host, where Figure 8 gives the number of active hosts for each protocol and dataset. The results are shown in Figure 9, where the label "smtp/10" shows the *smtp* connection rate divided by 10, in order that all three rates can be legibly shown on the same plot. The dashed lines on the plot show possible exponential fits. The fit to *smtp* per capita connection growth gives 29%/year, and the *ftp* fit, 37%/year. *Telnet*, on the other hand, does not show strong growth. Its fit corresponds to 5%/year.

We interpret these results as follows. Given that LBL has more than one computer per full-time staff member, we equate individual computers with individual users. With this equation, each active host corresponds to a user for whom part of their work patterns involves using the Internet for the service corresponding to the given protocol. (This argument is

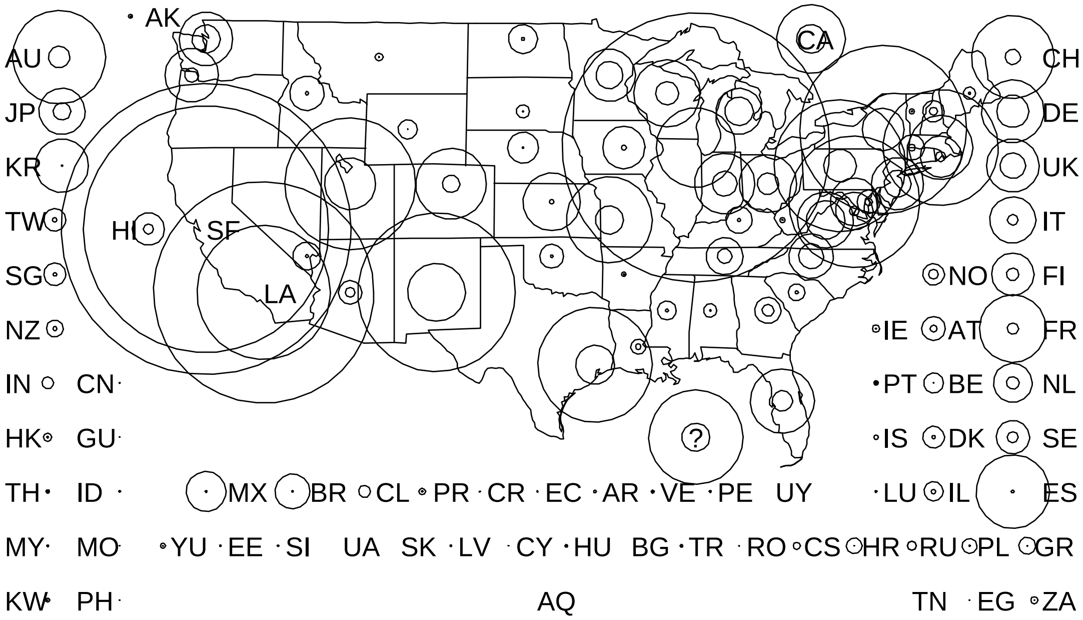

Figure 10: Geographical Distribution of LBL's Data Bytes

strongest for *telnet* and *ftp*, which involve the most local hosts, and weakest for *smtp*, which involves the fewest.) Then we interpret Figure 9 as showing how, on average, each user's work patterns varied over time. In particular, those users who use *telnet* do so at a steady rate, making on average one connection a day, while those using *ftp* and *smtp* are increasing their daily use rapidly. Similarly, *www* traffic has grown from an average of 1 conn./host/day in LBL-6, to 6 conn./host/day in LBL-8, indicating much greater individual use.

This finding suggests that even if the number of active hosts at a site stays constant, *ftp*, *smtp*, and *www* traffic is likely to keep growing exponentially.

# 7 Connection Geography

We conclude our study with a look at the geography of LBL's wide-area connections. For each remote host we attempted to identify (primarily using the *whois* server; see [HSF85]) the host's state or country. For each such region, we then computed the number of data bytes sent to or received from that region during the different datasets, and divided by the total number of bytes transmitted during the dataset, obtaining the proportion of the total traffic that involved each region.

Figure 10 shows the results. At each region on the map we have drawn two concentric circles. The area of each circle represents the fraction of LBL bytes that went to or came from the particular state or country encompassed by the circle. The larger circle shows that fraction's maximum value, and the smaller circle the minimum. For example, if a larger circle has twice the diameter (hence four times the area) as a smaller circle, then the greatest fraction of any dataset's bytes involving that region was four times the least fraction. If the pair of circles are close to the same size, then the fraction of bytes involving that region remained almost constant over the eight datasets.

In general, circles are either drawn centered at a region's most populous city, or the geographic center of the region when that results in less visual clutter (the state of New York, for example). Traffic to California is split into northern traffic (labeled "SF" for San Francisco) and southern traffic ("LA" for Los Angeles, though much of the traffic involves San Diego). LBL is sited in the "SF" region. As mentioned in Section 3, we exclude traffic to neighboring UCB except for *nntp* traffic.

Along the sides and bottom of the map we have drawn circles for the foreign countries with which LBL connected; these are marked with the country's two-letter ISO code. The region marked "?" corresponds to destinations which we either were unable to identify or could not pinpoint geographically (e.g., the `army.mil` domain).

Two observations are immediately apparent from the map. The first is that the bulk of LBL's traffic involves only a handful of regions, in line with the "favorite-site" effect discussed in Section 6 above. The second is that, discounting the volume of traffic, the geographical reach of LBL's connections is very wide. We discuss each of these in turn.

LBL's traffic mainly involves hosts in the southwestern and northeastern United States, with the largest portion remaining within California (and especially in the San Francisco region, where LBL is sited). From Figure 6 we know that *ftpdata* and *nntp* connections dominate the traffic by volume. LBL's primary *nntp* peers reside in Berkeley, San Diego, and Utah. Furthermore, researchers at LBL often collaborate with colleagues at other federal laboratories, particularly those in Illinois, New Mexico, Texas, and Switzerland (ISO code "CH").

We tested the data for discernible increasing or decreasing trends in each region's traffic fraction. No region had a consistent trend across all eight datasets or even across seven of the eight datasets (which we would expect to occur in any particular region by chance about 1% of time).

The variability shown in Figure 10, and especially the lack of consistent trends, argues for using considerable caution when interpreting a geographic snapshots of network traffic: the profile may well change a great deal over a relatively short period of time.

In addition to geography, we also investigated the distribution of data bytes to different top-level domains. The bulk of the traffic consistently involved `edu` sites (growing about 100%/year). The `gov` traffic was always second, but growing rapidly (155%/year), probably due to the collaborations with other federal laboratories discussed above. The third most popular destination was one of the foreign countries (which we aggregated into a single domain), growing at about 150%/year without appreciable slow-down.

One might expect the proportion of traffic to `com` sites to grow with time due to the increasing commercialization of the Internet. Our data suggests a very recent explosion in commercial use of the Internet: the first seven datasets show steady 72%/year growth of traffic to hosts in the `com` domain, but this traffic increased by a factor of *five* between LBL-7 and LBL-8.

The other top-level domains lagged these four considerably.

We finish our look at geography with a comment on geographic diversity. Of the 50 states in the U.S., we observed LBL Internet connections to every one. (Except for South Dakota and West Virginia, this diverse connectivity held as far back as LBL-3.) Wells reported 69 countries connected to the Internet as of May, 1994 [Wells94]; LBL had connections to 65 countries, illustrating very wide geographic use of the Internet.

# 8 Implications and Limitations

The main conclusions of our work are summarized in Table 1 above. The implications of our findings must all be tempered with the consideration that we studied the wide-area traffic of only a single site, with only three year's worth of data from which to infer trends. Clearly statistics from other sites are needed to gauge the generality of our results.

That a site with 30% annual growth in its computing facilities experienced substantially more rapid exponential growth in its wide-area traffic implies that network traffic grows at a significantly faster rate than growth in the number of hosts. To this end, statistics on network backbone growth in terms of total bytes transferred are an invaluable complement to connectivity-growth statistics.

The extremely rapid growth of new protocols such as *www*, *gopher*, and *X11* indicates that we should not rely on the logistic growth of older, more mature protocols as upper bounds on the wide-area network traffic a site will generate.

That the site's traffic profile was repeatedly skewed by inadvertent and casually-initiated TCP connections argues that sites would benefit considerably by monitoring their traffic profile.

Our findings that the number of computers participating in wide-area connections outpaces the site's overall host-growth imply that we should expect exponential growth to continue for a time after a site's host-growth tapers off. Furthermore, as per-capita use of some protocols appears to grow exponentially, even with a completely stable user community we would expect wide-area network use to continue growing exponentially for a while, only at a slower rate.

Finally, that a site's geographic traffic profile varies greatly over time implies that decisions regarding choosing backbone topologies (for wide-area network planners) or optimal service providers (for site planners) must be made based on multiple snapshots of the traffic profile.

# 9 Acknowledgments

This work benefited greatly from the advice and support of Van Jacobson and Domenico Ferrari. The LBL traces were gathered with the help of Craig Leres and Steve McCanne, and Craig was also very helpful in understanding various network phenomena.

I want to thank Jon Crowcroft, Domenico Ferrari, Van Jacobson, John Milburn, Keith Sklower, Jack Zelver, the reviewers, and particularly Sally Floyd for their comments on various aspects of this work, and ideas on how to present the results.

I also want to thank Rick Adams, Eric Aupperle, Vijay Gurbaxani, John Quarterman, and Don Wells for the help they provided in gathering network growth statistics.

This work was supported by the Director, Office of Energy Research, Scientific Computing Staff, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

# References

[Adams93] R. Adams, private communication, November, 1993.

[AML+93] F. Anklesaria, et. al, "The Internet Gopher Protocol", RFC 1436, Network Information Center, SRI International, Menlo Park, CA, 1993.

[BlCGP92] T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollermann, *World-Wide Web: The Information Universe*, Electronic Networking: Research, Applications, and Policy, 2(1), pp. 52-58, Spring, 1992.

[CR86] D. Cass and M. Rose, "ISO Transport Services on Top of the TCP", RFC 983, Network Information Center, SRI International, Menlo Park, CA, 1986.

[CPB93] K. Claffy, G. Polyzos, and H.W. Braun, "Traffic Characteristics of the T1 NSFNET Backbone", Proceedings of INFOCOM '93, San Francisco, March, 1993.

[DJCME92] P. Danzig, S. Jamin, R. Cáceres, D. Mitzel, and D. Estrin, *An Empirical Workload Model for Driving Wide-area TCP/IP Network Simulations*, Internetworking: Research and Experience, 3 (1), pp. 1-26, 1992.

[FJ93] S. Floyd and V. Jacobson, "The Synchronization of Periodic Routing Messages," SIGCOMM '93, pp. 33-44, September 1993.

[Gurbaxani90] V. Gurbaxani, *Diffusion in Computing Networks: The Case of Bitnet*, Communications of the ACM, 33(22), pp. 65-75, December, 1990.

[HSF85] K. Harrenstien, M. Stahl, and E. Feinler, "NICNAME/WHOIS", RFC 954, Network Information Center, SRI International, Menlo Park, CA, 1985.

[JLM89] V. Jacobson, C. Leres, and S. McCanne, *tcpdump*, available via anonymous ftp to ftp.ee.lbl.gov, June, 1989.

[Kleinrock76] L. Kleinrock, "Queueing Systems, Volume II: Computer Applications", John Wiley & Sons, 1976.

[Lottor92] M. Lottor, "Internet Growth (1981-1991)", RFC 1296, Network Information Center, SRI International, Menlo Park, CA, 1992.

[Lottor93] M. Lottor, SRI International, statistics available via anonymous ftp to ftp.nisc.sri.com, directory *pub/zone*; November, 1993.

[Merit94] Merit, Inc. Statistics available via anonymous ftp to nic.merit.edu, directory *nsfnet/statistics*; May, 1994.

[MJ93] S. McCanne and V. Jacobson, "The BSD Packet Filter: A New Architecture for User-level Packet Capture", Proceedings of the 1993 Winter USENIX Conference, San Diego, CA.

[Paxson93] V. Paxson, "Empirically-Derived Analytic Models of Wide-Area TCP Connections: Extended Report", technical report LBL-34086, Lawrence Berkeley Laboratory, May, 1993. Available via anonymous ftp to ftp.ee.lbl.gov, in the *papers/* subdirectory; retrieve WAN-TCP-models.prelim.1.ps.Z and WAN-TCP-models.prelim.2.ps.Z.

[Quarterman90] J. Quarterman, "The Matrix", Digital Press, 1990.

[Quarterman93] J. Quarterman, plots available via anonymous ftp to tic.com, directory *matrix/growth*; February, 1993.

[Rutkowski93] A. Rutkowski, *Internet Services: An Incredible Growing Cornucopia*, Internet Society NEWS, 2(2), pp. 19-22, Summer, 1993.

[SEKN92] M. Schwartz, A. Emtage, B. Kahle, and B. Neuman, *A Comparison of Internet Resource Discovery Approaches*, Computing Systems, 5(4), pp. 461-493, Fall, 1992.

[ST91] M. Schwartz and P. Tsirigotis, *Experience with a Semantically Cognizant Internet White Pages Directory Tool*, Internetworking: Research and Experience, 2 (1), pp. 23-50, 1991.

[Solensky92] F. Solensky, *The Growing Internet*, Connexions, 6(5), pp. 46-48, May, 1992.

[Terpstra93] M. Terpstra, statistics available via anonymous ftp to ftp.ripe.net, directory *ripe/hostcount*; November, 1993.

[WLC92] I. Wakeman, D. Lewis, and J. Crowcroft, "Traffic Analysis of Trans-Atlantic Traffic", Proceedings of INET'92, Kyoto, Japan, 1992.

[WRH92] C. Weider, J. Reynolds, and S. Heker, "Technical Overview of Directory Services Using the X.500 Protocol", RFC 1309, Network Information Center, SRI International, Menlo Park, CA, 1992.

[Wells94] D. Wells, "69 IP-connected countries now", USENET newsgroup comp.protocols.tcp-ip.domains, message DWELLS.94May6105756 @fits.cv.nrao.edu, May 6, 1994.