THE CURRENT STATE AND LIKELY EVOLUTION OF THE INTERNET

Andrew Odlyzko
AT&T Labs
Florham Park, NJ 07932, USA
amo@research.att.com
http://www.research.att.com/~amo

Abstract

Surprisingly little is known about the Internet. Even such basic facts as the size of the networks that make up the Internet or the amount of traffic they carry are not available.

This paper presents estimates of the main statistics about the size and growth of the Internet, as well as about utilization patterns. This data is then used to justify some speculative predictions about the likely evolution of data networks.

1. Introduction

This paper presents some of the highlights of the studies of data networks that are documented in [2, 5, 11, 12, 13] and in a few cases updates them. Much more detail about methodologies and results is available in those papers. This paper and those studies consider only high-level aggregate measurements of the Internet, and do not look at details of protocols, say.

There are many studies of the economics of the Internet. Most of them are listed in [9, 10, 14]. However, they are old (by Internet standards) and none of them answer such basic questions as how large the various parts of the Internet are, and how much they cost.

A key point of the investigation of [2, 5, 11, 12, 13] was the need to consider not just the public Internet, but the full universe of data networks and their role in the economy. For simplicity, only networks inside the U.S. were considered. Since costs of transmission are much lower in the U.S. than in most other countries, these networks are likely to reflect the behavior of the Internet in other parts of the world as costs come down.

Even in the restricted realm of data networks, the public Internet (those parts of the Internet accessible to general users) is only a fraction, although a noticeable and rapidly growing fraction, of the total system. Measuring networks by their maximal transmission capacity, it was estimated in [2] that at the end of 1997 in the U.S., the switched voice network was probably still the largest, but the private line networks were about as large, and the public Internet was considerably smaller. More recent updates of the estimates of [2], using the same methodology, show the following estimates for the end of 1998. (The bandwidth of data networks in the table is the effective bandwidth, as defined in [2], which is about half of the sum of bandwidths of all links. This measure was introduced to compensate for most packets traveling over about two links, as well as for data links being shorter than voice links. See [2] for the detailed justification.)

network	bandwidth (Gbps)
US voice	375
public Internet	150
other public data networks	80
private line	400

Thus looking just at the public Internet does not give a proper perspective on data networks, especially since utilization patterns of private networks are considerably different, as will be explained below.

Although data networks are about as large as the voice network in bandwidth, the voice network still dominates in carried load, and is likely to do so for a few more years. The traffic, measured in TB/month (terabytes per month), through various networks at the end of 1998 is estimated to have been (in another update of [2]):

network	traffic (TB/month)
US voice	43,000
public Internet	5,000 - 8,000
other public data networks	1000
private line	4,000 - 7,000

A comparison of the two tables above shows that there are substantial differences in utilization rates between the voice network and data networks. These differences can be used to infer what user preferences in data services are, and how much they are willing to pay. The basic argument (others are discussed later and in the papers mentioned before) is that low utilization rates show that what matters to users is the peak bandwidth, the ability to carry out transactions quickly, and not the

ability to send many bits. That is clearly what is driving the development of local area networks, and the evidence cited in this paper shows that most long distance data networks behave that way.

Section 2 is devoted to disproving a variety of common myths about the Internet. Section 3 presents some speculations about the evolution of the Internet.

2. Common wisdom or common misconceptions?

Much of the "folk knowledge" about the Internet is simply false. This section discusses the most important examples.

Traffic on the Internet is "only" doubling every year. Many press accounts, even in the professional data networking world, continue to claim that traffic on the Internet doubles every three to four months, corresponding to annual growth rates of 700% to 1,500%. The paper [2] showed that traffic on Internet backbones did grow about 1,000% in each of 1995 and 1996. We do have fairly reliable statistics for traffic at the end of 1994, when most of it was on the NSF backbone, and also for the end of 1996, when most is thought to have passed through the public peering points, for which data is available. Thus the high growth rates for 1995 and 1996 appear to be trustworthy. Less complete data appeared to show that by 1997 growth had slowed down to about 100% a year [2]. That is a remarkably high growth rate, but nowhere near as high as claimed in the popular press accounts. An update of [2] showed that this same growth rate of about 100% seemed to hold also in 1998. These estimates are not precise, and the true growth rate could be 80% or 120%, but it is almost certainly well below 200%, much less the "doubling every three months" that is sometimes cited.

Packet networks are not necessarily more efficient than the switched voice network. In general publications it is often asserted without qualification that packet networks are less expensive than the switched voice network. Some of the new packet-only carriers have been claiming that IP transport saves more than 90% over the cost of traditional switched networks. In particular, savings on transport costs are widely perceived as the main advantages of carrying voice over packet networks. On the other hand, when one considers existing corporate networks, and compares total costs and the volume of traffic, then it appears [12] that most corporations spend more on transferring large files over their internal IP networks than they would if they used modems over the public switched voice network. This is an astounding result, since modems use only a small fraction of the bandwidth of the digital channel that is provided for voice calls, and network costs of voice calls are small compared to the prices charged. The estimates for the cost (to corporations, which is not the same as the cost to the carriers) of transmitting a megabyte of data over various networks are estimated in [12] as follows:

network	dollars/MB
modem	0.25 - 0.50
private line	0.50 - 1.00
Frame Relay	0.30
Internet	0.04 - 0.15

This table suggests an obvious question: Why don't corporations junk their private networks and send data via modems over the public switched voice network? The answer is that the cost estimates of the table apply only to large file transfers, and do not take into account other factors, such as latency. As an example, a credit card authorization involves transfer of only a few hundred bytes, and so would cost far more over a modem than the table might suggest. It would also take far longer, tens of seconds instead of seconds, and thus lead to lower productivity of the sales force and customer dissatisfaction. There are thus unbeatable advantages to packet networks, but they are not in network costs, but in flexibility.

The above table raises another question, namely why don't corporations junk their private networks and send data via the Internet? This time the primary reason is the lack of security and high transmission quality on the Internet. Another reason is inertia, which is shown by the slow transition from private line networks to Frame Relay, which does provide the security and high transmission quality that the Internet lacks. Frame Relay is growing rapidly, at almost the rate of the Internet, but is not stopping private lines from continuing to grow.

The public Internet is still small relative to other data networks. Although it is the public Internet that has caught all the attention, it is still dwarfed in transmission capacity and especially in costs by the private line networks, as was shown in the tables in the Introduction. It is also far smaller than the switched voice network. However, it is growing much faster, about 100% a year, than either the voice network, which is growing at around 10% per year, or the private line networks, which are growing at around 20-30% per year. (See [2] for details. The information about the sizes of the private line networks is derived from published accounts by consulting companies that specialize in collecting such data, such as Vertical Systems. The estimates about the size of Internet backbones were assembled from a variety of sources, primarily network maps published by ISPs, usually accessible through the online Boardwatch directory at ⟨http://www.boardwatch.com⟩.) Therefore

if current growth rates continue, then in a few years the public Internet will be the dominant communication network, but it is not that yet.

Few data networks are congested. A surprising fact is that even though it provides high quality service, the switched voice network has considerably higher average utilization than any large collection of data networks. There is a general perception that the public Internet is hopelessly crowded, and even most network experts believe that private line networks are congested as well. Reality is different, as is shown in [11] and summarized in the table below. (The utilization rates in the table above, and elsewhere in this paper, refer to averages over a full week. Busy hour averages are higher, of course. For example, for private line networks, the busiest hour of a business day typically sees utilization of 15-25% of capacity, whereas for the voice network the corresponding figure is around 70%. However, it is long term averages that point out most drastically the different utilization patterns of the various networks and suggest ways to improve economics of data transport.)

network	utilization
local phone line	4%
U.S. long distance switched voice	33%
Internet backbones	10-15%
private line networks	3-5%
LANs	1%

The low utilization of data networks is a key finding that underlies most of the interpretations and speculations that follow later in this paper. This finding was extremely controversial when it was first publicized during the summer of 1998 in the preprint [11] (the first work to study this question systematically), and it is still not universally accepted. Particularly suspect was the claim that the Internet backbones were utilized at half or even one third the rate of the voice network. However, there is now more data available supporting these estimates. For example, AboveNet, a substantial ISP with a national backbone and even a trans-Atlantic link, makes detailed statistics for its network publicly available (at ⟨http://www.above.net/traffic⟩). In the first half of 1999 these statistics showed that the utilization of AboveNet's long-haul backbone was running around 16%.

The estimates of network utilization in [11] were based on extensive data for a variety of networks. Still, even those studies leave much to be desired. In particular, data about utilization of private line networks is scarce, although they form the bulk of all data networks. That is why the private line entry in the table spans nearly a factor of two. For details of the data, see [11].

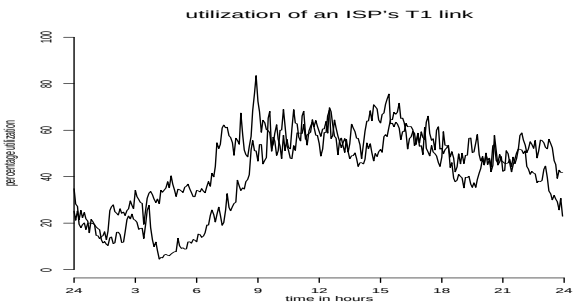

Figure 1: Traffic on an ISP's T1 line on Tuesdays of April 14 and 21, 1998. 5-minute averages.

Some parts of the Internet are highly congested, especially the public peering points, the NAPs and MAEs. Many university links to the public Internet are also heavily loaded, which may have persuaded generations of students that all networks are heavily utilized. However, the backbones of the Internet are relatively lightly loaded. The estimates of their utilization rates in [11] (based partially on estimates of sizes of various networks and the traffic they carry in [2]) are consistent with recent measurements which show that as long as transmission stays on a single backbone, latency and jitter are not a problem. What is congested are many of the feeder links to the backbones from smaller ISPs, especially those that aggregate modem traffic. Fig. 1 shows the traffic pattern on a T1 line (1.5 Mbps) belonging to an ISP. It runs at a high fraction of its capacity for large parts of the day, but still manages to provide relatively high quality service, with minor delays and packet losses, according to the network manager in charge of that link. (Average utilization is in the 40-45% range.) Other ISP links show even higher utilization, frequent saturation, and high packet loss rates. (Other examples of traffic patterns of ISPs, as well as those of other users, are in [11, 12, 13].)

As the tables in the Introduction show, most of the data transmission capacity is in private corporate networks. Their traffic patterns tend to be far different from those of the ISP line profiled in Fig. 1. Fig. 2 shows utilization of a corporate T1 line. The average utilization of this line is slightly under 1%. Comparing the graphs of Figures 1 and 2, it is easy to grasp that the performance of those two lines will be different, and that traffic control algorithms suitable for one might not fit the other one.

While most corporate networks are run at low average utilizations, there are many exceptions. The most

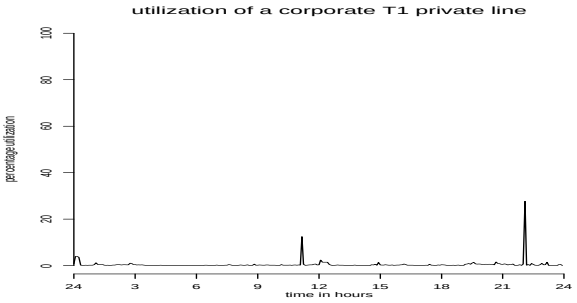

Figure 2: Traffic on a corporate T1 line in the continental U.S. during Thursday, May 28, 1998. 5-minute averages.

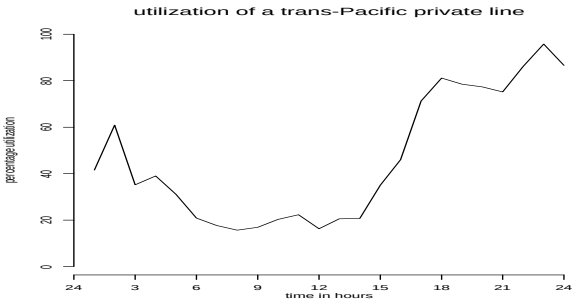

Figure 3: Traffic from the U.S. to the Far East on a corporate 128 Kbps line during a weekday. The peak traffic hours between 1800 and 2400 coincide with the busy hours in the Far East location. Hourly averages.

prominent are international lines, such as the one profiled in Fig. 3. The average traffic shown there is 59 Kbps, or 46% of capacity during the day that is profiled. On this particular link there is little traffic in the reverse direction, so average utilization of the entire line (which, as is always the case in current data network as a legacy of the switched voice network, consists of two one-directional links), during a business day is around 25%. Over a full week, average utilization might therefore be expected to be around 20%. However, according to the network managers in charge of the line, it does experience high packet loss rate (in excess of 25%) during peak traffic periods, and provides low quality transmission.

Congestion is not necessarily the biggest problem on the Internet. The "World Wide Wait" is often caused by problems other than lack of bandwidth. A study

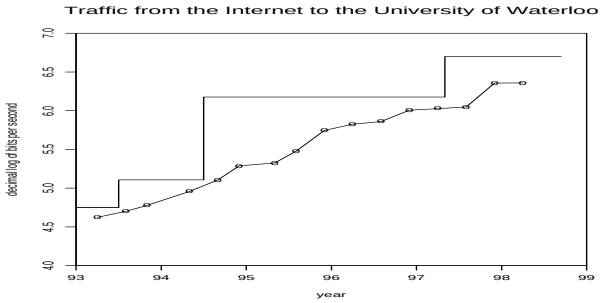

Figure 4: Traffic on the link from the public Internet to the University of Waterloo. The line with circles shows average traffic during the month of heaviest traffic in each school term. The step function is the full capacity of the link. By permission of University of Waterloo.

by Huitema [7] about accessing some popular servers showed that 20% were not reachable. Among the 80% that could be reached, 42% of the delays were caused by network transmission, with DNS accounting for 13% and servers for the remaining 45%. Further, there are some indications that in the last year, the performance of the backbones has improved, while servers are falling behind.

"The tragedy of the commons" may not be an insurmountable threat for the Internet. It is widely believed that queueing by congestion is how the Internet is run right now, and that this will not change until usage sensitive pricing is introduced, since demand, driven by flat rate pricing, is insatiable [6]. However, in a dynamic environment with growing bandwidth, this argument is questionable. In many cases, growth has been orderly. Fig. 4 shows the average traffic from the public Internet to the University of Waterloo. (See [2, 12] for more details.) Although the capacity of the link has had several sudden jumps, usage has grown at a pretty steady 100% a year. Similar steady growth rates have been seen in other networks, see [2]. Thus these networks have not in general had to cope with sudden surges in demand that saturated new capacity as soon as it became available. Even when such surges materialized (as they did at the University of Waterloo when student dorms were hooked up to the campus Ethernet), they were contained by simple local measures, primarily quotas on traffic to individual PCs.

Not only is growth of data traffic steady, actual traffic is generally predictable once it is sufficiently aggregated. Several of the graphs in this paper, as well as many of those in [11, 12] combine displays of traffic for

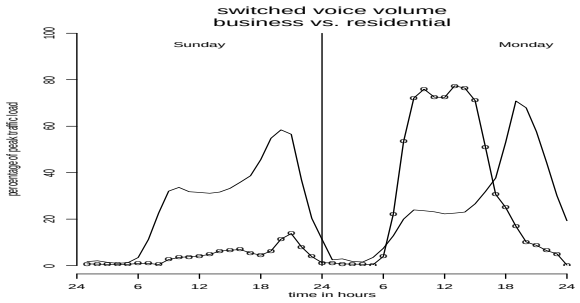

Figure 5: Residential (thin line) and business (line with circles) voice traffic on U.S. long distance switched voice networks, as percentage of peak traffic on those networks.

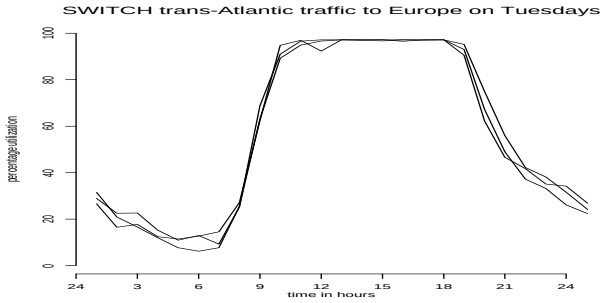

Figure 6: Traffic on the 8 Mbps link from the U.S. to SWITCH, the Swiss academic and research network, during Tuesdays of February 3, 10, and 17, 1998. Hourly averages, Swiss time. By permission of SWITCH.

several days. It is noteworthy that the traffic patterns are generally consistent from week to week, with Monday through Thursday usually behaving the same, and Friday, Saturday, and Sunday each having its own particular load graph. This is the same behavior that has been observed on the switched voice network.

There are many inefficiencies in data networks that are not being exploited. Most attention is currently devoted to Quality of Service (QoS) measures [4], but without providing quantitative estimates of how much such measures will save, or will improve the quality of transmission. However, there are many other steps that can be taken to provide a better Internet, steps whose benefits can often be quantified much more easily and reliably. For example, costs could be lowered if utilization were increased by combining corporate private line traffic on public networks, using Virtual Private Networks (VPNs). A major reason for the high utilization rate of the switched voice network is that it carries traffic from both business and residential customers, and those two classes of users have complementary traffic patterns, as is shown in Fig. 5.

Historically there has been considerable asymmetry in public Internet traffic between the U.S. and Europe and Asia, with the U.S. sending more bytes than it receives. This asymmetry has been increasing in the last couple of years, so that on many links the ratio is 2:1 or even 3:1. (For example, on the British JANET network, in March 1997, 3.73 TB were received from the U.S, and 2.95 were sent there, for a ratio of of 1.26. In March 1999, the corresponding figures were 19.52 and 9.51, for a ratio of 2.05. On the Swiss SWITCH network, during the month ending on Feb. 4, 1999, the corresponding traffic figures were 3.34 and 1.29, for a ratio of 2.59.) Hence one could obtain much better capacity utilization by building transmission systems that do not have the symmetric links that are a heritage of the switched voice network. The gain from doing this is easy to quantify, unlike the potential gains from implementing many of the current QoS measures, for which there seem to be no hard numerical figures.

The "bursty nature of data traffic" is not the culprit behind low utilization rates of data networks. Data traffic does not smooth out as well as switched voice traffic, and it shows long range dependence [3, 8]. However, that does not mean that high utilization cannot be achieved. Figures 3 and 6 show that it can. In Fig. 6, we see essentially full utilization over 9 hours during the business day, and in Fig. 3, for a much smaller link, more than 80% utilization over a comparable period. (The "goodput," or measure of traffic that end users care about, ignoring the retransmissions, is presumably much lower, but no estimates for it are available.) Most of data transport uses TCP, which fills available bandwidth and can produce high load factors. Thus low utilization has to come from a different source.

The paper [11] identified a variety of reasons why utilization rates of data networks are likely to lag behind those of the voice network. These reasons include rapid growth rate, asymmetry in data traffic, and the much lower prices per unit of bandwidth of higher capacity links. These reasons by themselves explain most of the difference in utilization patterns of ISP networks and the voice network. Private line networks have the additional disadvantage that they carry traffic primarily during the business day, and thus lose the advantage of having complementary traffic patterns that help keep ISP and voice system pipes full. Still, that does not

fully explain the low utilization rates of private line networks.

In some cases there is a clear rationale for the design of data networks. When large files are to be backed up to an off-site facility, the networks are sized appropriately to that task, with some margin of safety, and show moderate utilization rates. In other cases, say in online transaction processing, there are stringent requirements for how long a transaction can take, and networks are designed accordingly, usually resulting in low utilization. In most cases, though, there is no clear rationale, and designers use a variety of "rules-of-thumb," such as not allowing the utilization of a T1 to exceed 50% over more than a certain fraction of 5-minute intervals during a business day. Ultimately such rules appear to come from subjective judgements of the end-users. Looking at utilization rates and utilization patterns, it appears that the main driving force in the development of data networks is the desire for low transaction latency. (This is my interpretation of the data, and it has to be said that it is not universally accepted.) Customers do not care about networks as such, only about applications. Only a few people are consciously aware of what they are doing with data networks. As an example of such a person, the manager of a branch lab of a major software producer, who has a private line from that lab to company headquarters, said that

> I see peak bandwidth as the basic commodity I buy. ... When we had a 256Kb data line it was too slow (it interfered with productivity). With a T1 line, no one has complained. I guess our T1 line is less than 1% utilized. ... I would not go for a T3 line (it would not improve our productivity) but I would not cut back on the T1 line.

High bandwidth can to some extent compensate for high packet latency (cf. [1]), since it is the time for the total transaction (such as a Web page download) that matters to the user, not the time that the first packet makes it through. (In cases of extreme latency, such as satellite channels, protocols that spoof TCP by sending false acknowledgements to the server from a gateway are often employed to allow the full bandwidth to be utilized.) The main point, though, is that high bandwidth is absolutely essential for low transaction latency. If a 5 MB PowerPoint presentation has to be transmitted from a telecommuter's home to her office, it will take over 20 minutes over a 56 Kbps modem (which can transmit upstream at only around 30 Kbps), but in favorable conditions under a minute over a good ADSL connection. That private line networks in the continental U.S. have low utilization rates shows that the desire for low transaction latency is the driver when costs are not too high. That similar private lines across the oceans are heavily utilized, with high packet loss rates and similar impairments, shows that when costs are very high, the end-user desire for low transaction latency is subordinated to the need to lower costs through high utilization.

3. The future of the Internet

In this section I speculate about the future of the Internet. These speculations are based on the facts presented in previous sections and conclusions that I drew from those facts.

In the debates about the future of the Internet, there are arguments for preserving a single best-effort service class without state inside the network, and with low utilization providing high quality of service for all traffic. Such arguments have often been supported by citations of progress in optical transmission, which promises much lower costs for data links. The counterargument has typically been that no matter how low the cost, there would always be some cost, and so the service providers would have an incentive to maximize utilization and therefore would have congested links. Another, related counterargument has been that data networks suffer from "the tragedy of the commons," and no matter how much transmission capacity is built, it will quickly fill up (as happens almost universally with roads) [6].

The observations of the preceding section provide strong evidence in favor of the hypothesis that one can build a best-effort stateless backbone network that will offer high quality transport to all traffic primarily through low utilization. The argument is more subtle, but also more convincing, than simply saying that prices of transmission will come down, and therefore we will be able to afford larger pipes. The point is that, as was discussed at the end of the preceding section, what people care about is not transmitting bits, but transmitting them quickly, to achieve low transaction latency. Therefore, if prices of data networks decrease as fast as technological progress suggests they should, data networks will likely evolve in the same direction that LANs and computers have, namely towards low utilization. Note that the counterarguments cited at the beginning of this section (about data pipes filling up) ought to apply to LANs and PCs, yet both of them are very lightly utilized. Although PCs are not free, new 400 MHz Pentium II machines are being bought even for secretaries. Yet old 486 machines could in principle do the job. Those fast new PCs are purchased for their peak performance, to load

word processors or to recalculate spreadsheets rapidly. Their average utilization is immaterial. Similarly, 10 Mbps LANs are being replaced by 100 Mbps ones, and 100 Mbps LANs are beginning to be replaced with gigabit LANs not because the older networks could not carry the traffic that is offered to them, but because these older slower networks have high transaction latency. The evidence about low utilization of private line networks in the continental U.S. shows that people want the same properties of their long distance data links that they demand from their LANs and their PCs. Further, in the majority of cases they are already able to buy this performance by obtaining high bandwidth links that they use at low rates. Therefore it seems likely that as prices of data links decrease, the Internet will evolve towards lightly utilized links. (For more details on this argument, and further ones, see [13].)

A caveat that has to be offered is that the conclusion about the feasibility and desirability of a single best-effort service class is based on two key assumption. One is that prices of data transmission will decrease in line with progress in photonics. (As is documented in [2], but is not widely known, prices in the U.S. that are paid by corporate network managers and ISPs that do not own their own physical network did decrease rapidly during the 1980s, but have been climbing since 1992. Recently there have been some signs of a change, and definite declines have been reported in other countries that had not experienced the North American declines of the 1980s, but there is no general trend of decreasing prices yet.) The other assumption is that traffic on the Internet will continue to be dominated by transactions such as Web-surfing and file transfers, and not by real-time video and audio. The argument for the second assumption is that while there is likely to be extensive video and audio traffic, it will be in the form of file transfers (such as MP3 ones) for later playback on a variety of information appliances, and not in streaming form. However, this is definitely a hypothesis. For arguments supporting this hypothesis, see [13]. One of those arguments comes from observing the development of LANs. Those already are moving towards speeds of 100 Mbps and above, which are far more than enough to accommodate streaming media. This shows that when prices are sufficiently low, the desire for low transaction latency does produce high bandwidths all by itself.

References

[1] J. P. Cavanagh, *Frame Relay Applications: Business and Technical Case Studies,* Morgan Kaufman, 1998.

[2] K. G. Coffman and A. M. Odlyzko, The size and growth rate of the Internet. *First Monday,* Oct. 1998, ⟨http://firstmonday.org/⟩. Also available at ⟨http://www.research.att.com/∼amo⟩.

[3] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, The changing nature of network traffic: Scaling phenomena, *Computer Communication Review,* 28, no. 2 (April 1998), pp. 5–29. Available at ⟨http://www.research.att.com/∼agilbert⟩.

[4] P. Ferguson and G. Huston, *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* Wiley, 1998.

[5] P. C. Fishburn and A. M. Odlyzko, Dynamic behavior of differential pricing and Quality of Service options for the Internet, pp. 128–139 in *Proc. First Intern. Conf. on Information and Computation Economies (ICE-98)*, ACM Press, 1998. Available at ⟨http://www.research.att.com/∼amo⟩.

[6] A. Gupta, D. O. Stahl, and A. B. Whinston, The Internet: A future tragedy of the commons?, available at ⟨http://cism.bus.utexas.edu/res/wp.html⟩.

[7] C. Huitema, The required steps towards high quality Internet services, unpublished Bellcore report, 1997.

[8] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Networking* 2 (1994), 1-15.

[9] J. MacKie-Mason, Telecom Information Resources on the Internet, Web site with links to online sources, ⟨http://china.si.umich.edu/telecom/telecom-info.html⟩.

[10] L. W. McKnight and J. P. Bailey, eds., *Internet Economics,* MIT Press, 1997. Preliminary version of many papers available in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[11] A. M. Odlyzko, Data networks are lightly utilized, and will stay that way. Available at ⟨http://www.research.att.com/∼amo⟩.

[12] A. M. Odlyzko, The economics of the Internet: Utility, utilization, pricing, and Quality of Service. Available at ⟨http://www.research.att.com/∼amo⟩.

[13] A. M. Odlyzko, The Internet and other networks: Utilization rates and their implications. Available at ⟨http://www.research.att.com/∼amo⟩.

[14] H. R. Varian, The economics of the Internet, information goods, intellectual property and related issues, reference Web pages with links, ⟨http://www.sims.berkeley.edu/resources/infoecon/⟩.