



CHAPTER

3 Broadband Microfoundations: The Need for Traffic Data

Steven Bauer, David Clark, William Lehr

<https://doi.org/10.5422/fordham/9780823251834.003.0003> Pages 51–68

Published: June 2013

Abstract

This chapter explains the importance of traffic data in providing a richer picture of the overall state of broadband networks. It begins with an overview of what traffic data is generally available in networks, how this data may prove important in answering questions that are relevant to the entire community of stakeholders, and why collecting the data and making it more generally available is challenging. It then considers how data focused on broadband availability and on adoption metrics is becoming less informative as broadband penetration rises, and that network traffic data will be central to understanding and answering many of the questions relating to the health of the broadband access market. To answer these questions, it will be helpful to learn more about the distribution of usage across the user population, the characteristics of users that participate during peak periods of network congestion, and the variance in usage and how it differs by type of user. The chapter concludes with a discussion of open research questions and some of the interesting research efforts that have been initiated in recent years.

Keywords: [traffic data](#), [broadband networks](#), [broadband availability](#), [broadband adoption](#), [metrics](#), [broadband penetration](#), [network traffic](#), [broadband access](#), [network congestion](#)

Subject: [Museums, Libraries, and Information Sciences](#)

To date, most of the empirical effort to understand broadband service markets has focused on availability and adoption metrics and data. Data of this sort is indeed valuable when the dominant policy questions concern penetration and uptake. However, as broadband availability and penetration saturate, such data will become less informative. The next set of questions, both for service providers and regulators, will center on the continued health of the broadband access market: levels of investment, the competitive landscape, the evolving definition of broadband, the degree of neutrality in consumer access, and the nature of interconnection among providers. Our position is that network traffic data will be central to understanding and answering many of these questions. To answer these sorts of questions it will be helpful to know such things as the distribution of usage across the user population, the characteristics of users that participate during peak periods of network congestion, and the variance in usage and how it differs by type of user. This data will help inform forecasts of capacity/infrastructure investment needs (e.g., how much bandwidth does a subscriber need? How much sharing is feasible at which points in the network?), to understand ISP costs, and to assess network management practices (e.g., traffic engineering). Better traffic data will provide insights into consumer adoption decisions and the evaluation of product offerings (e.g., how important are peak rates versus average data rates?).

The Internet is often compared to the network of highways, streets, and roads that make up the transportation system. Both are vital infrastructures that provide businesses with access to materials and markets, and provide people with access to goods, services, recreation, jobs, and each other. For transportation networks, it is generally recognized that traffic data (i.e., the volume of traffic, congestion information, incident reports, and so on) is as important to understanding the state of the network as is information about where the roads or links actually are. The same is true for the Internet.

In transportation networks, traffic data is valuable over both short time scales (e.g., allowing real-time traffic management to reroute commuters around a rush hour accident) and over longer time scales (e.g., for planning maintenance cycles and capacity expansion investments). During periods of congestion,¹ traffic data and real-time traffic management via lights, tolls, and special commuter lanes has proved especially important in enabling more efficient utilization of the existing transportation infrastructure. Improving the efficiency of the existing infrastructure delivers benefits in the form of reduced commute times (contributing directly to labor productivity), improved safety, and reduced pollutant emissions through intelligent traffic management policies.²

On the Internet, traffic data is similarly important to network operations. Over short time scales ranging from less than a second to hours or days, traffic data is an input into systems (both automated and human-centered) that make routing decisions (e.g., balancing loads across different network links), identify suspected or actual security or transmission failures, and implement traffic management policies.³ Over longer time scales measuring months or years, traffic data is vital to capacity planning and provisioning, allowing capacity to be efficiently installed in advance of demand, thereby better accommodating future traffic growth without congestion-related disruptions. Thus traffic data is essential to almost all the practical dimensions of network management and to the political, regulatory, and theoretical questions of what constitutes good, acceptable, or socially desirable network management.

p. 53

While traffic conditions on the highway and roadways can be observed externally (via both technical sensors and human observations), information about Internet traffic and the level of congestion of the different autonomous networks that collectively compose the Internet is limited. While individual network operators generally have a good idea about the state of their own networks, outside stakeholders have little visibility into the state of traffic on networks. Networks can be probed and tested by outside observers to derive some measurements, but the scope and confidence of such measurements is limited compared to the accuracy and breadth of information available to network operators.

The majority of users have very little visibility or understanding of what is happening to their traffic once it enters a network. Without better visibility, it is not surprising that there are widely diverging opinions about the true state of networks. For example, what are the congestion and utilization levels now and in the predictable future? What are the underlying cost structures for carrying traffic and expanding capacity? Or, what are the effects of different traffic management policies?⁴

The problem is that this limited visibility by outside stakeholders into the traffic and congestion state of networks makes it hard to have confidence in the regulatory and investment decisions that affect such networks. On the one hand, traffic management policies that are efficient and “fair”⁵ could be disrupted or private investments in expanding capacity could be deterred. On the other hand, network operators could be exploiting their control to thwart or discourage disruptive new innovations and competitors (either intentionally or accidentally).

Furthermore, in contrast to our transportation grids where most of the physical infrastructure (roads, terminals, bridges) are publicly funded and managed, most of the physical infrastructure that composes the Internet is investor-funded and privately managed. We rely chiefly on profit-motivated firms and market competition to direct resources to their best uses for the collective benefit of society and the economy. For markets to work efficiently, they depend on the public availability of relevant market information to allow buyers and sellers to formulate their strategic decision-making. For example, market participants need to understand what they are purchasing or selling, between whom the exchanges are to occur, and the timing and terms for market transactions. Markets generally work best when they are lightly regulated, so ensuring that the appropriate information is produced by the market process presents an interesting challenge for institutional design and incentive compatibility.⁶

Our thesis is that better visibility by outside stakeholders into the traffic data of networks is required to improve the regulatory processes, investment/market decision-making, and technical research. More generally, better visibility of what traffic looks like in networks will promote understanding and trust between what ultimately has to be a cooperative community of interconnecting and communicating parties.

p. 54

At least some network operators are interested in sharing their internal data in order to facilitate this process (see below). The challenges to making this happen are multidisciplinary: engaging aspects that are technical (how to sample or share the potentially terabyte-sized data sets), analytic (how to compare and combine data generated by different measurement processes), ↪ policy-oriented (how to preserve the privacy of individual subscribers), and business strategy related (how to protect competing providers' business interests). Addressing these problems while still producing data capable of providing useful insights into the important questions noted above is nontrivial and requires a multifaceted and process-oriented approach that is capable of evolving as the Internet evolves.

In the following sections we give an overview of what traffic data is generally available in networks, how this data may prove important in answering questions that are relevant to the entire community of stakeholders, and why collecting the data and making it more generally available is challenging. We conclude with a discussion of open research questions and a brief overview of some of the interesting research efforts that have been initiated in recent years.

Traffic Data

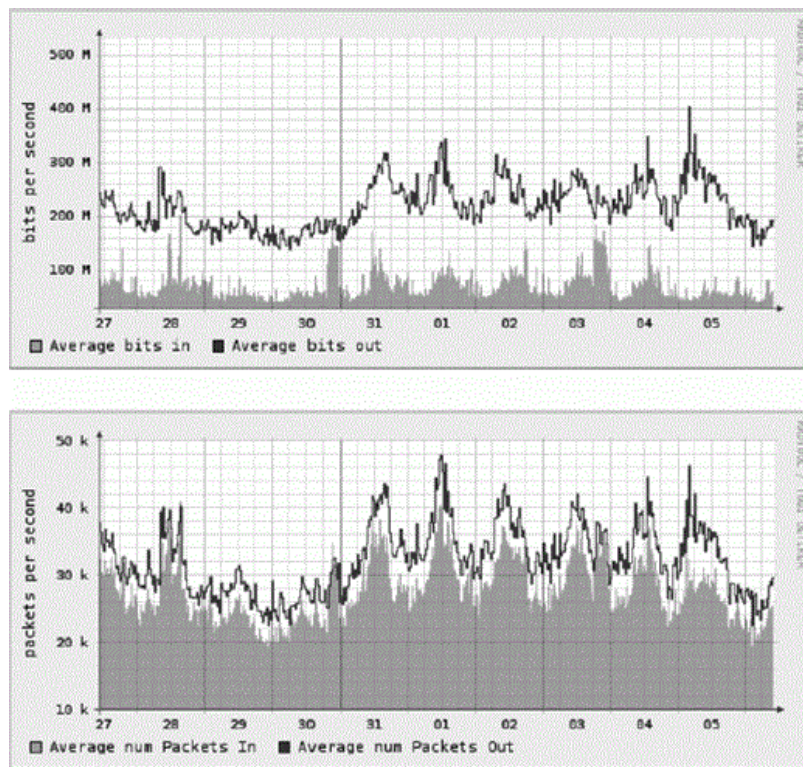
The amount of information that *could* be collected by network operators from their networks is enormous. Individual network elements—which include routers, switches, servers, caches, and subscriber modems—can report hundreds of different statistics. With hundreds of thousands of elements in a network, and millions of subscriber lines, the volume of potential data is enormous. For example, one network operator we spoke with indicated that the total volume of data records could exceed three hundred terabytes of data a year. Collecting and transporting the raw data in real time to the network operations center where it can be processed, analyzed, and managed presents a difficult challenge that incurs significant operational costs. Determining what data to archive and how to compress or summarize the data and manage access presents complex statistical, logistical, and policy challenges.

In spite of the costs, network operators do systematically collect real-time traffic data because it is essential for successful network operation. The data is an input into strategic and operational decision-making across virtually all ISP functions. The data informs decisions about the capacity of internal links, routing policies, security policies, and interconnection contracting. It is used for high availability and disaster recovery planning, for financial projections, employee evaluations, technical strategy discussions, and sales and marketing. In larger network operations, there are specialized departments focused on managing the collection and analysis of network traffic data, and the sharing of relevant portions and views of the data across the organization.

p. 55

One of the most important uses for the traffic data, after monitoring the health of the existing network, is capacity planning. This is accomplished by studying the utilization of network links averaged over multiple time intervals. ↪ The utilization data of a link is collected from a router using a protocol such as SNMP (Simple Network Management Protocol). Figure 3-1 shows utilization graphs from one of the gigabit Ethernet links connecting our lab to the main MIT campus network. Our lab sends more traffic (top line) than it receives (bottom solid color) because we host a number of popular sites including mirrors of software distributions. One can see in the graph the diurnal variations in traffic displayed. On the top are the average bits per second and on the bottom are the average packets per second. Both are potentially important statistics as a router can be congested because of the volume of data (each packet carrying the maximum amount of data) or the number of packets (each packet could have little data but there could be hundreds of thousands of packets). Congestion in most networks today is more likely related to excess volume than excess packets.

Figure 3-1.



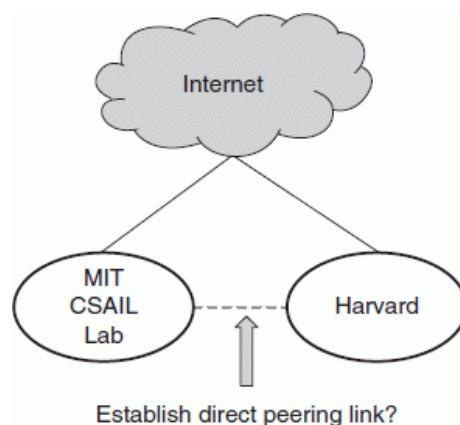
Measurement of bits per second.

For this particular link at MIT, there is no particular evidence of any persistent performance or congestion problems. While we don't display it here, the long-term trends on this link also don't suggest congestion or

p. 56 performance problems in the near future as there isn't significant growth in the aggregate traffic levels. However, if there were hints of impending traffic congestion, other forms of data would be instrumental in analyzing the causes and planning the course of action, and to understand trends, it would be necessary to have time-series data documenting utilization across time.

A network operator at our lab might first look at flow-level details using data such as Netflow records.⁷ These records provide a way of looking inside the aggregate flow to better understand what combinations of edge sources and destinations (forming a traffic matrix) are actually communicating. Such data is essential to understanding whether peering or upstream connections should be modified. For instance, this data might indicate that our lab sent and received a significant amount of traffic to and from the Harvard campus. Therefore we might be able to reduce the utilization on the loaded link by establishing a separate direct peering connection to Harvard, thereby offloading some traffic to the new link (see Figure 3-2).

Figure 3-2.



Traffic matrixes derived from Netflow style data is one way of determining where peering links should be established.

Another way of reducing link utilization is to constrain the top contributors to traffic on a link. Table 3-1 is a list of top traffic contributors, again derived from Netflow data. At our lab this data is monitored primarily to identify anomalous sources of traffic such as hosts that have been infected and are unwittingly serving as

bots or data depots for hackers. But given that it is a shared research network, disputes can arise as to what constitutes acceptable use of network resources. These tools provide a way of objectively identifying “hot spots” and measuring the impact of different experiments, websites, and uses.

Table 3-1 Top Contributors to Traffic on the Csai Network for one Period in September

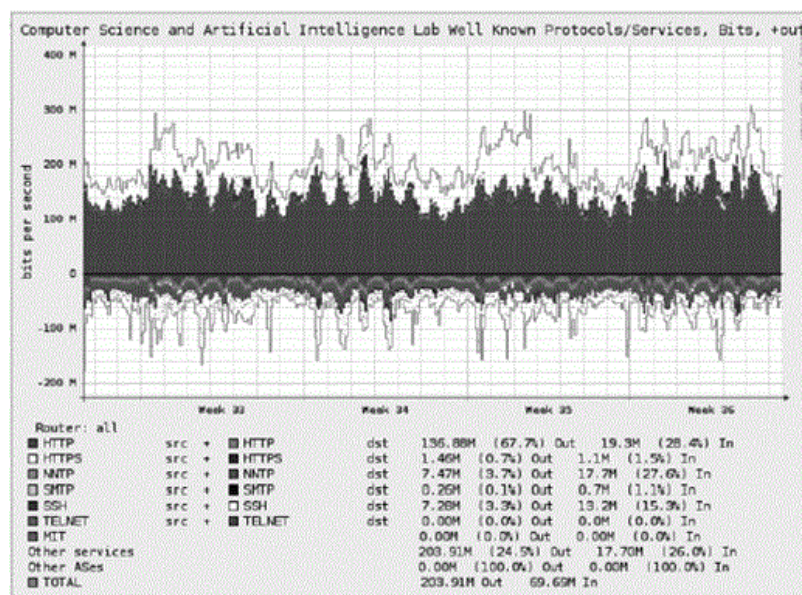
Rank	Address	Bit/sec In	Bits/sec Out	Pkts/sec In	Pkts/sec Out	Flows/sec In	Flows/sec Out
1	carver.debian.org 128.31.0.50 (1 sample)	60.0 M	1.4 M	6.3 k	3.2 k	223.3 m	233.3 m
2	infmite-state.csail.mit.edu 128.30.24.177 (1 sample)	48.2 M	1.3 M	4.1 k	2.1 k	3.3 m	3.3 m
3	rore.debian.org 128.31.0.49 (1 sample)	38.1 M	735.2 k	4.1 k	463.3	29.7	28.9
4	mosdef.w3.org 128.30.55.83 (4 samples)	35.4 M	573.8 k	3.1 k	1.3 k	10.8 m	10.0 m
5	newsswitch.csail.mit.edu 128.30.2.35 (37 samples)	18.1 M	7.6 M	1.9 k	1.7 k	241.5 m	240.9 m
6	thursday.csail.mit.edu 128.30.100.224 (1 sample)	10.2 M	200.7 k	897.9	463.8	13.3 m	23.3 m
7	30-7-158.wireless.csail.mit.edu 128.30.7.158(3 samples)	4.6 M	71.2 k	390.7	184.9	250.0 m	248.9 m
8	planetlab3.csail.mit.edu 128.31.1.13 (37 samples)	4.1 M	3.5 M	956.2	957.6	85.9	93.5
9	mdemaine.csail.mit.edu 128.30.48.115 (2 samples)	3.4 M	57.3 k	300.5	109.3	1.4	1.4
10	xyz.csail.mit.edu 128.31.0.28 (19 samples)	3.3 M	3.7 M	504.7	514.5	1.7	2.1

p. 57 Another way in which detailed traffic data is employed is to examine what protocols and applications are being used on a network. This data is significant ↴

both to identify anomalies (a significant rise or drop in any category might indicate a problem) and to understand and predict future traffic growth. (Figure 3-3 shows a sample of the protocols in use on our lab network.) Particularly as new video-centric applications and services become more ↴

p. 58

Figure 3-3.



Top contributors to traffic on CSAILs lab network for one period in September 2009.

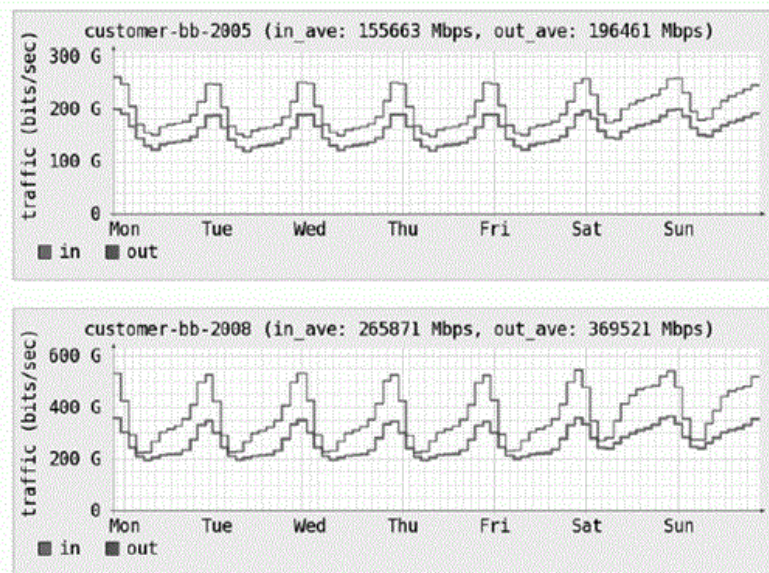
popular, monitoring their adoption will be crucial to capacity planning.⁸ Many of the emerging applications transmit their data over random ports or standard web ports (thereby mixing in with other types of web traffic) so Netflow data records may become less useful for monitoring the adoption of “new” applications over time. Other tools and measurement devices—often referred to as Deep Packet Inspection or DPI—enable more detailed traffic analysis on a per flow or per packet basis. These techniques seek to classify traffic flows by looking at other information both within the packets and other predictable signatures such as the pattern of communication (bytes transmitted during the initial connection handshake, and so on).⁹

While individual ISPs collect such data on their networks, they have little insight into the detailed traffic patterns on other networks, even ones they may be directly connected to. In spite of the need for such data to monitor the macro-economic health and direction of the broadband marketplace, such data is not readily available publicly. A notable exception is the excellent collaborative research project among ISPs that has been underway in Japan since 2004.¹⁰ That project represents the most advanced publicly reported broadband data project undertaken to date—seven large ISPs, carrying roughly 40 percent of Japanese traffic, contributed summary data on traffic characteristics at least twice yearly since 2004. This data offered a compelling picture of the growth and distribution of broadband traffic as experienced in Japan.

p. 59

While there are many different interesting details that emerged from their work, we highlight some here to give concrete examples of how traffic data connects to important macro-economic issues. Unsurprisingly, the transition to broadband has fundamentally changed traffic patterns on the Internet. The effects of this transformation are sometimes obvious, but also sometimes surprising. For many years, the peak usage periods of access networks (which generally serve both residential and commercial customers) were during the business day. However, for at least the last several years, the peak usage hours of many access networks are in the evening, roughly between 9 p.m. and 11 p.m. (see Figure 3-4). This is important for understanding the economics of networks as the previously off-peak residential customers used to more easily “fit” in the pipes that had been provisioned for the peak-using commercial users. Now, however, the usage patterns of the residential customers are often driving the provisioning decisions of network providers.¹¹ This has obvious implications for cost sharing and service pricing.

Figure 3-4.



Residential broadband traffic in May 2005 (top) and May 2008 (bottom) as measured in Japan.

Source: Cho, 2008.

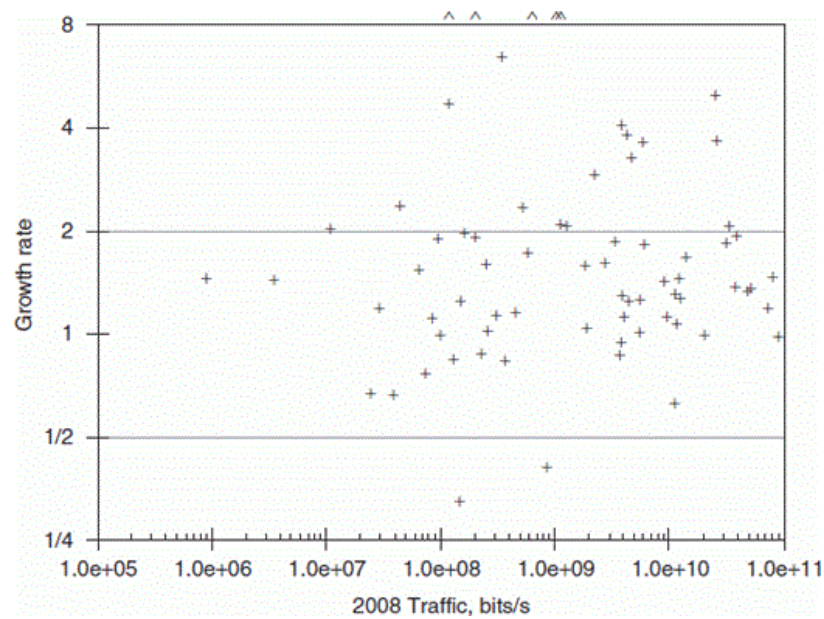
The Minnesota Internet Traffic Studies (MINTS), run by Andrew Odlyzko at the University of Minnesota, has been monitoring traffic growth

levels on networks for a number of years.¹² The traffic data in his study is derived from publicly available data sources such as peering points and sites, such as universities, that post information about their traffic. Most of his data comes directly from Multi Router Traffic Grapher (MRTG) and Round Robin Database (RRD) graphs (very similar to the previous figures in this chapter). The raw data used to generate the graphs would be even more informative and presumably preferred by most analysts, but most sites do not make it

p. 60

available. The MINTS data shows that the aggregate level of traffic continues to grow at double-digit rates, recently averaging around 50–60 percent compound annual growth rate (CAGR) per year (see Figure 3–5).¹³ While these growth rates are impressive, they are substantially lower than the rates widely cited in the trade press over the years.¹⁴

Figure 3-5.

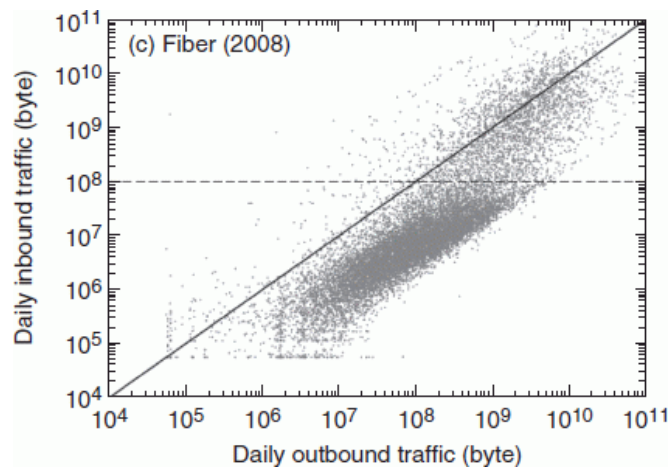


2008 traffic growth rates from publicly observed sites in the MINTS traffic study.

Kenjiro Cho et al. (2006) used the Japanese ISP data to investigate how the mix of applications on broadband networks is changing. Addressing one of the most significant questions for the near-term traffic growth—the macro-level impact of video—Cho noted that “the current traffic is heavily affected by an eruption of peer-to-peer applications but the crust underneath is also slowly rising with video and other rich media content. The crustal movement is slow at the macro level so that it is unlikely to cause a major quake in the near future.”¹⁵ This is a good metaphor as the increasingly

p. 61 popular video traffic does not pose an imminent threat to the stability of the Internet, but the growth in video traffic will be significant, eventually fundamentally reshaping the traffic mix on broadband networks. This will have unmistakable economic impacts on regulatory policy, innovation in new applications and services, competition, and the value chain of network vendors and suppliers.

The final question on which existing public data sheds some light is the distribution of traffic among subscribers on a network. This is significant because networks are shared resources where not all traffic demands can necessarily be simultaneously satisfied. So there is a very basic question as to what constitutes fair sharing of a network. Users are sometimes categorized as exhibiting “heavy” versus “regular” or “light” usage patterns. The relationship between the aggregate volumes of traffic a subscriber sends or receives and their contribution to congestion (in terms of causing packets to be dropped) is not always clear. It is possible that a “heavy user” does not disproportionately contribute to either packet dropping congestion or to usage during the aggregate peaks on a network. What is clear though is that there are very large differences in the volume of traffic sent and received by different subscribers. While most users may download less than two gigabytes of traffic in a month, the top users on a system can easily exceed one hundred gigabytes. Figure 3–6 displays the average *daily* inbound and outbound traffic per user on a fiber network in Japan measured over a week in 2008. Each dot represents one user.



Correlation of daily inbound and outbound traffic volumes per user in one Japanese metropolitan prefecture for a fiber optic network in 2008. Each dot above the dashed line represents users that sent more than 100 megabytes of traffic in a day.

Source: Cho, 2008.

p. 62 As peak rates increase, and hence the possibility for sending and receiving ever larger amounts of traffic grows, there exists the potential for an increasing divergence between the volumes of traffic that different segments of the market send and receive.¹⁶ This is not problematic in and of itself. A challenge will arise, however, if these very different usage patterns are associated with different underlying cost structures either in terms of the congestion they contribute to or in terms of the variable costs their usage incurs (e.g., transit charges from an upstream network provider).

The Importance of Traffic Data

The previous section provides just a sample of the extensive history in the networking community on research detailing the technical behavior of networks. Indeed, the properties of individual links, paths, hosts, and networks have been extensively analyzed. While these measurements have served the purpose for which they were designed, connecting these technical details and data to inform the economic, regulatory, and policy challenges of networking is a relatively new challenge.¹⁷

What are missing in most regions of the world are collections of data and measurements that provide a richer picture of the overall state of networks. As demonstrated, this is data that broadband providers routinely collect and analyze in their individual network operations centers, but is rarely understood or shared with the wider community, including other operators. By aggregating views of multiple individual networks, a picture of the issues, opportunities, and problems confronting both individual networks and the collection of networks that comprise the Internet¹⁸ can be developed while still protecting the confidentiality of individual network operators and subscribers.

In particular we see this data as important to establishing traffic trends, growth, and characterizations at both the aggregate and subscriber level; vital inputs into a data-driven discussion of network management practices; promoting public and industry awareness of the challenges, successes, and opportunities in the broadband marketplace; and assisting in diagnosing and understanding traffic problems and phenomena.

Broadband traffic characterization. Data about broadband and network traffic is needed to develop representative aggregate and subscriber traffic models that are used to analyze and forecast market trends and plan network provisioning and management. While aggregate growth statistics indicating the total volume of Internet traffic or subscribership are clearly important and regularly cited in company annual reports, municipal broadband plans, policy debates, research papers, and the popular press, more detailed and less aggregated data are needed to understand the composition of the aggregates, to identify local phenomena, and to discern the drivers and relationships among the subcomponents. Data on top-line growth alone is not adequate to address questions about the changing mix of applications (e.g., peer-to-peer versus streaming video), differences in platform technologies (e.g., cable modem versus DSL versus wireless), or changes over time (in response to changing technology, network architecture, or the industry ecosystem). Data to allow the decomposition of aggregate growth are needed for the development of rich

p. 63

future scenarios and to support flexible “what-if” analyses. User traffic data is needed to understand “within” and “across” user traffic distributions (e.g., how do subscriber usage patterns vary across subscribers and across time?). When linked to cost and revenue data, representative traffic data underlies a fuller understanding of broadband economics.

Traffic diagnosis. Better traffic data will enable the analysis of significant traffic events. A number of organizations currently produce analyses based upon their view of both public and private data. These include analyses of the effects of the de-peering incidents,¹⁹ significant cable cuts,²⁰ routing incidents,²¹ major media events,²² and security incidents.²³ Each of these provides important lessons learned in terms of understanding the actual and potential effects of the incidents and also learning how they might be prevented in the future. More traffic data would provide a richer picture of the effects of these incidents on communications, business, and users.

Traffic management. Representative traffic samples and broadband traffic models will prove useful in enhancing simulations and in testing network management approaches, including congestion management strategies.

This is particularly important now since access providers have met with opposition, from a mix of stakeholders, to the deployment of network devices that implement a variety of congestion management policies.²⁴ These policies often change the network resource allocations that would result from the distributed actions of host applications and TCP stacks. While the result is certainly different than what would occur without these devices, it is not de facto unfair or welfare reducing. However, the wider Internet community might regard it as unfair or inefficient depending upon the policies that are implemented.²⁵

Promoting public and industry awareness of the challenges, successes, and opportunities in broadband. There is a lack of awareness across the Internet value chain and within the wider community of the challenges posed for infrastructure investment, especially in last-mile networks, from Internet traffic growth. In 2004, as part of the Broadband Working Group in the MIT Communications Futures Program (a collaborative effort with academic and industry partners from across the broadband value chain) we examined what we termed the *broadband incentive problem*—the challenge of incentivizing ISPs to continue investing in expanded capacity in the face of rising traffic-related costs.²⁶ As household subscribership approaches saturation, the growth in access revenues priced at a flat monthly rate per subscriber line will slow, but aggregate traffic will continue to grow. Reduced investment by ISPs in expanding network capacity poses a threat to innovative, high-bandwidth uses of the Internet. How best to resolve this quandary is a challenge for the entire Internet value chain and will likely require a mix of new investment, new usage and pricing models, and better network management.

p. 64

Evaluating the economic health of the broadband marketplace is also important. There is a growing research literature documenting the economic benefits of the Internet and broadband for employment and productivity.²⁷ More granular data about subscriber usage patterns would help improve these studies and offer insight into how best to promote universal adoption and how best to target public broadband funds.²⁸

Public traffic data also would offer a perspective on where the opportunities in broadband are. In the future, understanding where broadband service is available and which households are subscribers will become less interesting relative to questions about how broadband is being used to support novel applications directed at improving education, health care, business processes, entertainment, and communication. Which regions are leading and lagging in the adoption of these innovations will be of general interest.²⁹

Challenges

In this section we discuss some of the challenges of collecting an appropriate multi-ISP traffic data set. One of the primary challenges is that the data requirements evolve over time as the measurement infrastructure changes (both in terms of measurement locations and methodology), the questions asked of the data change (requiring more granular or detailed data on a particular topic), and legal and regulatory obligations are modified (changing what can or must be collected). Thus the institutional frameworks put in place to gather data must be flexible and able to accommodate changes. This is nontrivial because forging even temporary agreement on a methodology requires the assent of the technical, legal, and management teams of all participating organizations.³⁰

Technical challenges. All the typical technical challenges associated with data collection arise: missing data, spurious data, missing metadata, and ambiguous fields. Data can and is commonly lost as systems are moved, upgraded, and reconfigured. If a data collection process for a network temporarily fails, it is often impossible to restore past data.³¹

p. 65 Varying network measurement methodologies are common over time, across ISPs and measurement equipment providers, and even within a single provider's network because the provider may have a mix of vendor equipment and legacy systems. The precise location of traffic probes in a network determines what traffic is measured. In the case of analysis boxes (or DPI)³² that identify applications and protocols, the choice of equipment vendor and the rules in effect at any point in time (i.e., what measurement options are set and the current generation of vendor software) have a considerable impact on how traffic is classified (e.g., how much traffic may be classified as "other"). Because traffic classification techniques differ across equipment vendors, one could expect different traffic classification results even for an identical stream of traffic. While we are not aware of any systematic study of the differences, the network operators we have spoken to indicate that such differences are common.

The sheer size of the data sets can also present challenges. Depending on the ISP, the data sets may range from small comma-separated data files of less than one megabyte in size to specialized databases that collect hundreds of terabytes of data a year. Large data sets often dictate that sampling procedures be employed otherwise even basic queries can take hours to run. One network operator we spoke with indicated that, in their initial data collection setup, running a database query at the same time as data was being collected was impossible.

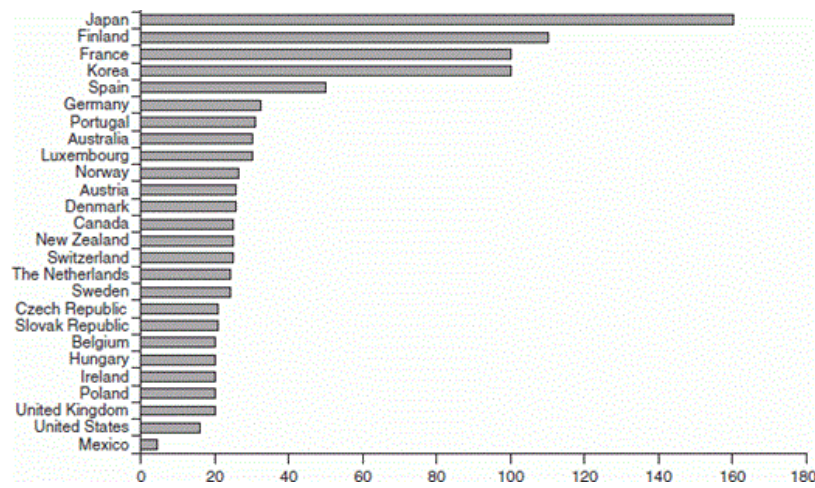
Analytic challenges. An area of particular interest is how to match traffic characterizations with other types of data in ways that allow analysts to better understand aggregate and per-user behavior while protecting against ex post-user identification (a challenge we discuss below). There is a great deal of information that would be desirable to collect and compare but that, in practice, is challenging to acquire. For instance, to better understand the drivers of user behavior, it would be desirable to understand what other services (telephony, video, premium video, and so on) a subscriber takes, the advertised service characteristics (peak rate, service pricing, and so on) of each subscriber, subscription timing (when was service first initiated, when changed, when terminated), geographic location data about the subscribers, and other types of demographic data.³³

At least in some providers' networks, it is hard to bring together service plan information with usage data. Not only are the databases physically separate, they also reside in separate organizational units within the business. Even the internal analysis teams are stymied at times when they seek to match usage and service description data. It is also challenging to answer some analytic questions for technical reasons. If one wanted to analyze how traffic demands shifted immediately following a capacity upgrade, it is difficult in practice to identify the precise timing for when the upgrade took effect for an individual subscriber. Just because a subscriber has been authorized to utilize higher peak sending and receiving rates, the subscriber may not have rebooted their modem to pick up the new settings and hence would still be running with older and slower rates until they do.

Even once data is collected, analysis is complicated because there are no generally accepted "correct" metrics for many of the questions that come up in discussion of traffic data. In a related paper,³⁴ we presented several different definitions for congestion and recounted some of the intellectual history in the evolution of thinking about congestion to suggest the importance of this complexity. Each metric has implications and is important in particular contexts, but can be misleading if not understood properly.

As another example, consider the Organization for Economic Cooperation and Development (OECD) report that documents the "fastest advertised broadband speeds" per country in 2008 (see Figure 3-7). What is advertised is technically very different in different countries, rendering cross-country comparisons of even something as seemingly straightforward as peak advertised rates difficult. For example, for similar technologies or services, the rates that are advertised in Japan and the United States are different. For example, the Japanese advertise a maximum peak rate of 160 Mbps that is not achievable in practice.³⁵

Figure 3-7.



OECD data from September 2008 on the fastest advertised broadband speeds using cable (in Mbps).

Legal challenges. The privacy implications of measuring individual subscriber behaviors must be carefully managed. There are obvious concerns that detailed information may enable socially undesirable forms of discrimination. Thus, there is a growing awareness that Internet traffic data needs to be managed so as to respect and protect subscriber privacy. Today, there are

p. 67 no clear or universally accepted norms or rules for protecting user data on the Internet. This is an active and important area of ongoing research.³⁶

Even in the absence of consensus on norms or rules,³⁷ ISPs have brand images to protect from perceptions that they may have inadequately protected subscriber privacy or misused subscriber data (regardless of whether any such perceptions are well-founded).³⁸ Thus, addressing concerns about protecting individual subscriber data are very important in generating support for the collection and sharing of appropriate public data. One of the benefits of pooling data from multiple ISPs is that it provides additional options for preserving provider and subscriber confidentiality, while enabling sufficiently rigorous and detailed sampling to obtain statistically accurate traffic characterizations.

For academic researchers, many universities, including MIT, require that all affiliated personnel that are engaged in research involving human subjects submit their proposal to an institutional review board (IRB)³⁹ that has the responsibility to confirm that the research is in compliance with federal regulations designed to protect human subjects from harm that may arise as a consequence of the proposed research. Risks to individual privacy are one of the potential harms that the IRB process is intended to address. While the original focus of these rules was on humans engaged in medical research, prompted by several well-known cases of abuse,⁴⁰ the IRB process has now been extended to all research involving human subjects. This process provides an additional layer of protection to ensure adequate privacy protection. There are a variety of techniques including sampling design and anonymization that may be used to ensure that the data that is collected does not include any personally identifiable information (PII).

Business challenges. As noted earlier, the efficiency of markets depends on the availability of adequate information to key stakeholders (buyers and sellers). There is a rich economics literature documenting the importance of private and asymmetric information, and its potential to effect the allocation of resources and profits.⁴¹ The fact that better traffic data may make the Internet ecosystem more competitive and efficient means that such data is inherently strategic. Better traffic data may allow an ISP to better plan its investments and target its service offerings to capture market share from other providers.

We believe efforts to collect data would be most successful if the data is voluntarily supplied. While the data *could* be compelled, using strong regulation to collect or force disclosure of the data would introduce regulatory costs (e.g., direct overhead as well as distorting incentives) and rigidities (technology evolves faster than regulations). If successful mechanisms can be crafted to make sharing the data incentive compatible, we believe the result will be to get more granular, timely, and better data publicly available sooner and at a lower total cost to stakeholders.⁴²

p. 68 There is also a free-rider problem that will need to be addressed. Even if one accepts the value of better public information on traffic, most folks would be happier if they can derive those benefits without having

to pay the costs of making the data available. While this will pose a challenge, it is hardly a new challenge or one limited to the problem of Internet traffic data, and so there are a host of well-known approaches for addressing this challenge. We expect that industry associations, consortia, and standardization bodies may play useful roles in figuring out how to resolve these issues.


Conclusion

In the initial phase of broadband Internet access, the focus of policy-makers and many researchers has been on ensuring universal availability and adoption of broadband service. As broadband subscribership saturates, broadband infrastructure continues to evolve (e.g., toward much higher potential peak rates, toward mobile broadband, and so on), and the applications enabled by broadband become more widely relied upon (e.g., interactive rich multimedia applications), questions about how broadband is being used will be increasingly of interest. To properly address such questions, we will need much better insight into Internet traffic (its growth, statistical characterization, drivers, and so on) over both short (operational) and longer (investment) time frames.

We have argued in this chapter that traffic data will be central to monitoring and resolving the inevitable tussles of this next stage of development. Given that such data is not publicly visible, the cooperation of network operators is essential. We are optimistic that the challenges of sharing traffic data can be addressed. The technical community (including academics, operators, vendors, and interested individuals) has a long history of collaborating through institutions such as the IETF,⁴³ NANOG (and its equivalents in other regions),⁴⁴ and other forums. A similar cooperative capacity can be developed which produces data about traffic on the Internet.

While we have a clear understanding of why such traffic data is important now, we also recognize the importance of collecting data in anticipation of future use. In “Looking Over the Fence at Networks: A Neighbor’s View of Networking Research” it was noted that “good data outlives bad theory.”⁴⁵ Data can be useful to later generations of researchers in ways not yet understood. The report noted the heavy dependence of the scientific community’s knowledge and understanding of climate change on a record of atmospheric carbon dioxide measurements that Charles David Keeling started collecting on Mauna Loa in 1957. An analogous historical data set of traffic data for the Internet might be similarly important for future networking research providing a baseline for evaluating the large-scale impact of both evolutionary and revolutionary changes in the Internet.

Notes

1. For a discussion of congestion in the Internet, see our companion paper: S. Bauer, D. Clark, and W. Lehr, “The Evolution of Internet Congestion,” prepared for the 37th Research Conference on Communication, Information and Internet Policy, Arlington, VA, September 2009.
2. D. Jonsson, S. Berglund, P. Almström, and S. Algers, “The Usefulness of Transport Models in Swedish Planning Practice,” *Transport Reviews: A Transnational Transdisciplinary Journal* 31 (2011): 251–265.
3. For an introduction to network management, see M. Subramanian, *Network Management: Principles and Practice* (Reading, MA: Addison Wesley, 1999).
4. Over 10,000 comments were filed in the FCC’s proceeding (07-52) regarding Comcast’s management of peer-to-peer traffic (see http://fjallfoss.fcc.gov/prod/ecfs/comsrch_v2.cgi).
5. We put “fair” in quotes because there is an on-going debate in the technical community as to what constitutes fair allocations of network traffic. See for example B. Briscoe, “Flow Rate Fairness: Dismantling a Religion,” *SIGCOMM Computer Communication Review* 37 (2007): 63–74; S. Floyd and M. Allman, “RFC 5290: Comments on the Usefulness of Simple Best-Effort Traffic,” Network Working Group, Internet Engineering Task Force, July 2008, <http://www.faqs.org/rfcs/rfc5290.html>.

6. For example, since public funds are used to build roads and bridges, and there is a strong interest in maintaining transparency in government processes, the need to collect and publish relevant information and statistics is well established and (relatively) straightforward. In contrast, where private investment is involved, ensuring adequate public disclosure of relevant information is more complex.
7. Netflow is the common name for this record type, but it has been standardized now as Internet Protocol Flow Information eXport (see <http://en.wikipedia.org/wiki/IPFIX>).
8. Different types of traffic have different profiles in terms of upstream/downstream bit rates, tolerance for delay, jitter, or bit error losses, and amenability to being multicast. Knowing the mix of applications facilitates planning to ensure an appropriate quality of service for the applications that are expected.

9. These techniques are imperfect because the traffic signatures change as new applications are introduced and because they are based on sampling techniques that are subject to stochastic measurement error.
10. K. Cho, K. Fukuda, H. Esaki, and A. Kato, "Observing Slow Crustal Movement in Residential User Traffic," in Proceedings of the 2008 ACM CoNEXT Conference, Madrid, Spain, December 2008; K. Cho, K. Fukuda, H. Esaki, and A. Kato, "The Impact and Implications of the Growth in Residential User-to-User Traffic," in Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, September 2006.
11. Cho et al. (2008).
12. See <http://www.dtc.umn.edu/mints/home.php>.
13. See Minnesota Internet Traffic Studies (MINTS) at <http://www.dtc.umn.edu/mints/home.php>, for data on traffic growth rates. ↗
14. For example, see "Net Traffic Doubling Every Six Months," a report from August 2001, http://www.theregister.co.uk/2001/08/17/net_traffic_doubling_every_six/.
15. Cho, et al. (2008).
16. In addition to distinct differences in the usage patterns of different types of users, there may be different numbers of each type; and they may be distributed differently across a network in ways that may be related to what they are doing (e.g., different on-net/off-net patterns) with resulting implications for aggregate traffic flows.
17. This is not to say there are not a number of researchers addressing this challenge. For example, in addition to the work by Odlyzko/MINTS and Cho/Japanese ISPs, see the work of Caida at http://www.caida.org/publications/papers/2009/aims_report/aims_report.xml#topten. We are aware of a number of other projects and suspect there are many more we are unaware of being undertaken at universities and in industry labs (e.g., Cable Labs, AT&T Labs) across the United States and abroad.
18. Most subscriber traffic is a mix of on-net (i.e., traffic that originates and terminates on the access ISPs network) and off-net (i.e., traffic that either originates or terminates on another ISP's network). Because ISPs generally lack detailed insight into traffic conditions on the networks of other ISPs, better pooled traffic data may allow ISPs a more complete understanding of the factors driving local phenomena on their own networks (e.g., separating local from general trends) as well as conditions in the wider Internet.
19. "Internet Captivity and the De-peering Menace," <http://www.renesys.com/tech/presentations/pdf/nanog-45-Internet-Peering.pdf>. ↗
20. "Deja Vu All Over Again: Cables Cut in the Mediterranean," <http://www.renesys.com/blog/2008/12/deja-vu-all-over-again-cables.shtml#more>. ↗
21. "The Day the YouTube Died: What Happened and What We Might Do About It," <http://www.renesys.com/tech/presentations/pdf/nanog43-hijack.pdf>. ↗
22. Akamai's "Net Usage Index for News" enables users to monitor global news consumption 24×7 , seeing in real time the impact of current events on online media consumption. <http://www.akamai.com/html/technology/nui/news/index.html> ↗
23. Conficker/Conflicker/Downadup worm as seen from the UCSD Network Telescope <http://www.caida.org/research/security/ms08-067/conficker.xml>. ↗
24. For example, regulatory authorities have initiated proceedings to examine ISP traffic management practices (e.g., for Canada see <http://www.crtc.gc.ca/ENG/archive/2008/pt2008-19.htm>, November 2008; and in the United States see http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-08-92A1.pdf, January 2008).
25. For further discussion of why these issues are contentious, see Bauer, Clark, and Lehr (2009).
26. Broadband Working Group, "Broadband Incentive Problem," a white paper by the MIT Communications Futures Program, September 2005, http://cfp.mit.edu/publications/CFP_Papers/Incentive_Whitepaper_09-28-05.pdf. ↗
27. See for example H. Varian, R. Litan, A. Elder, and J. Shutter, "The Net Impact Study: The Projected Economic Benefits of the Internet in the United States, United Kingdom, France, and Germany," research report, funding support from Cisco Systems is acknowledged, January 2002; W. Lehr, C. Osorio, S. Gillett, and M. Sirbu, "Measuring Broadband's Economic Impact," paper prepared at the Telecommunications Policy Research Conference, Arlington, VA, September 2005; S. Greenstein and R. McDevitt, "The Broadband Bonus: Accounting for Broadband Internet's Impact on U.S. GDP," white paper, Technology Policy Institute, Washington, D.C., January 2009, <http://www.techpolicyinstitute.org/files/greenstein-broadbandbonus1.pdf>; M. Dutz, J. Orszag, and R. Willig, "The Substantial Consumer Benefits of Broadband Connectivity for U.S. Households," *Compass Lexecon*, A Study Commissioned by the Internet Innovation Institute, July 2009.
28. For example, the American Recovery and Reinvestment Act of 2009 (Pub. L. 111-5, 123 Stat. 115, 2009) has targeted US\$7.2 billion in public funding for the promotion of broadband. Most of the data that is available regarding usage patterns is highly aggregated over geographic regions, time periods, users, and applications in ways that obscures many potentially interesting details.
29. See <http://www.connectivityscorecard.org/images/uploads/media/TheConnectivityReport2009.pdf>.
30. In Cho's Japanese study, the challenges were described as mainly political not technical. See Cho et al. (2008); Cho et al. (2006).
31. To economize on data storage costs, raw data is summarized in real time.
32. DPI refers to Deep Packet Inspection. Equipment and software are available from a variety of vendors that allows detailed inspection of the contents of packets, not just the header-information which must be examined in order to forward the packets toward their destination. Looking inside the packets for additional information about what the application is or other traffic identifier information is referred to as DPI.
33. This is information that is not available in the Japanese studies described above.
34. Bauer, Clark, and Lehr (2009).

35. See Figure 3-8, noting in particular Japan, which offers 160 Mbps broadband. This particular service is offered at 6,000 yen (\$60, according to a 2009 *New York Times* blog, <http://bits.blogs.nytimes.com/2009/04/03/the-cost-to-offer-the-worlds-fastest-broadband-20-per-home/>) and is a DOCSIS 3.0 service offered over a cable network. What is interesting is that if four channels are bonded together the maximum synchronization speed of a subscriber's modem and the CMTS is around 170 Mbps but the maximum usable speed is 152 Mbps (i.e., less than the advertised speed).
36. For a sample, see J. Camp, *Trust and Risk in Internet Commerce* (Cambridge, MA: MIT Press, 2001), Wik-Consult/Rand Europe/CLIP/CRID/GLOCOM, *Comparisons of Privacy and Trust Policies in the Area of Electronic Communications*, Final Report, report prepared for the European Commission, July 2007, http://ec.europa.eu/information_society/policy/ecomm/doc/library/ext_studies/privacy_trust_policies/final_report_20_07_07_pdf.pdf; P. Ohm, "The Rise and Fall of Invasive ISP Surveillance," *University of Illinois Law Review* 2009, 1417. Links to current policy debates and further research are available at Electronic Privacy Information Center (<http://epic.org/>); privacy.org (<http://privacy.org/>); and Electronic Frontier Foundation (<http://www.eff.org/issues/privacy>).
37. ISPs do face regulatory rules on the disclosure of subscriber data, but these rules vary by context and are not comprehensive.
38. In today's Internet blogosphere rumors of bad behavior can be damaging.
39. The IRB is a review board established by MIT and with similar boards at other universities with representation from across the community.
40. These included the Milgram Obedience to Authority experiments of 1988 and the Public Health Service Syphilis Study (1932–1972), more commonly referred to as the "Tuskegee Syphilis" experiments.
41. See for example J. Tirole, *The Theory of Industrial Organization* (Cambridge, MA: MIT Press, 1988).
42. Government-mandated data collection of detailed data comes with strong nondisclosure obligations.
43. Internet Engineering Task Force, <http://www.ietf.org/>.
44. The North American Operators Group, <http://www.merit.edu/nanog>.
45. C STB, *Looking Over the Fence at Networks: A Neighbor's View of Networking Research* (National Academy Press, 2001).