Broadband microfoundations: the need for traffic data

Steven Bauer[1], David Clark[2], William Lehr[3]

Massachusetts Institute of Technology

Abstract

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?).

1. Introduction

The Internet is often compared to the network of highways, streets, and roads that make up the transportation system. Both are vital infrastructure 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, etc.) 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.

[1] Corresponding author: bauer@mit.edu, tel: 617-869-5208.

[2] ddc@csail.mit.edu tel: 617-253-6002

[3] wlehr@mit.edu, tel: 617-258-0630

In transportation networks, traffic data is valuable over both short time scales (e.g., allowing real time traffic management to re-route 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,[4] 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. (See Figure 1 for a picture of traffic conditions in the Boston area. Traffic data in the map above is derived from roadway sensors and location updates from individual's cell phones.[5])

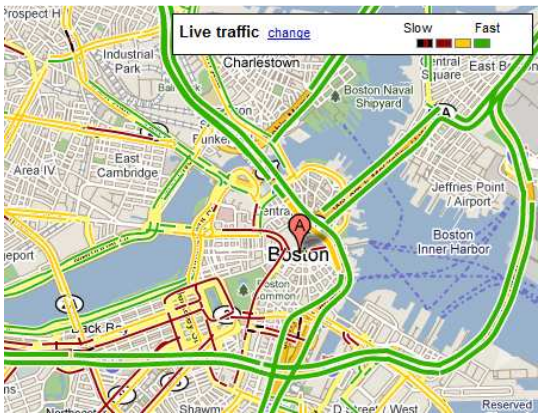

Figure 1: Traffic data enables drivers to avoid congested roadways and plan trips.

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 incipient or actual security or transmission failures, and implement traffic management policies.[6] A number of these traffic management policies, which are particularly significant for both regulatory and user experience reasons, operate at the level of individual subscriber flows. These include techniques such as 1) volume caps that limit the total volume of traffic offered by a subscriber over different durations of times and in the upstream and or downstream directions, 2) prioritizing subscriber or application traffic based upon factors such as the amount of traffic sent during congested periods or assumptions regarding what traffic subscribers would prefer to be prioritized (such as voice traffic over bulk transfers) and 3) rate

[4] For a discussion of congestion in the Internet, see our companion paper, Bauer, Clark, Lehr (2009).

[5] See http://news.cnet.com/8301-30684_3-10317223-265.html for a description of the cell phone derived traffic data. We find this particularly interesting because collaboration of multiple parties and technologies is essential to producing an aggregate picture of the current state of traffic.

[6] See Subramanian (1999) for an introduction to network management.

limiting traffic classes, such as peer-to-peer traffic, that are believed to significantly contribute to congestion. 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/theoretical questions of what constitutes good/acceptable/socially desirable network management.

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 congestion state of the different autonomous networks which collectively composite 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 breath 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. Their experience is analogous to driving with black painted windows, slowing down and speeding up based only upon whether or not they are bumping into anything in front of them. They have little understanding of whether the bumps are result of congestion with other traffic, speed bumps put in place by the network operators, or other bottlenecks such as the limited capacity of the destination they are trying to reach. Without better visibility, it is not surprising that there are widely diverging opinions about the true state of networks (i.e. what are the congestion and utilization levels now and in the predicable future, what are the underlying cost structures for carrying traffic and expanding capacity, and what are the effects of different traffic management policies).[7]

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. There is a risk of making decisions that have undesirable or unexpected results and a symmetric risk of failing to make decisions that would have been beneficial. On one hand, traffic management policies that are efficient and 'fair'[8] 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).

---

[7] See, for instance, the widely varying opinions offered in comments to the FCC in its proceedings (07-52) which followed from Comcast's management of peer-to-peer traffic. A search of http://fjallfoss.fcc.gov/prod/ecfs/comsrch_v2.cgi for comments in this proceeding turns up over 10,000 results of opinions offered by individuals, companies, service and application providers, and academics.

[8] We quote 'fair' here because there is debate about what actually constitutes fair allocations of network traffic. See for instance Briscoe (2007) and Floyd and Allman (2008).

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 (who/what to purchase from/sell to? How much/when to purchase/sell?). 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.[9]

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. Without a clear understanding of what traffic is doing now, and where it might be headed in the future, making good decisions is considerably more difficult. More generally, better visibility of the traffic state of networks will promote understanding and trust between what ultimately has to be a cooperative community of interconnecting and communicating parties.

At least some network operators are interested in sharing their internal data to facilitate this process.[10] 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 non-trivial and requires a multi-faceted 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 some of the interesting research efforts that have been initiated in recent years.

---

[9] For example, since public funds are used to build roads and bridges, and there is a strong interest in maintaining a transparency in government processes, the need to collect and publish relevant information and statistics is well established and (relatively) straight-forward. In contrast, where private investment is involved, ensuring adequate public disclosure of relevant information is more complex. We have detailed accounting rules, intellectual property protection, and rely on markets (with advertising and a multiplicity of supplier options) to generate adequate public information.

[10] See Section 5.

2. Traffic data

The amount of information that *could* be collected by network operators from their networks is enormous. Each individual network element (routers, switches, servers, caches, subscriber modems, etc.) can report hundreds of different statistics, values, and events. With hundreds to 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 total volume of data records could exceed 300 terabytes of data a year.[11] 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/summarize the data and manage access present complex statistical, logistical, and policy challenges. One operator's joking comment was "my job would be a lot easier if no one actually wanted to do anything with this data."

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.

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 some time intervals. The utilization data of a link is collected from a router using a protocol such as SNMP.[12] The following in Figure 2 are 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. On the left is displayed the average bits per second and on the right is 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.

[11] In this particular case, these were IPDR data records which provide per subscriber usage information. With a large enough user base and a collection frequency of multiple times per hour, the volume of data understandably grows to be large quickly.

[12] SNMP stands for Simple Network Management Protocol.

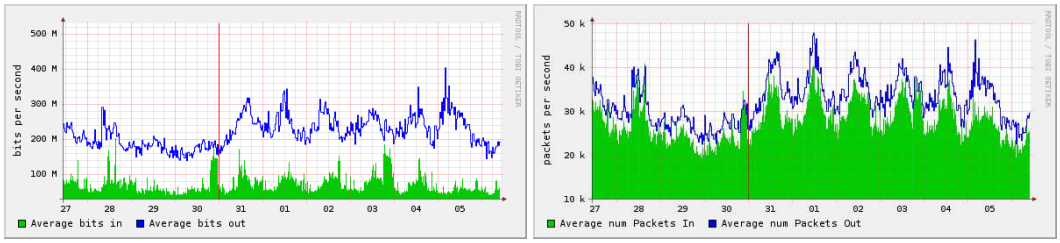

Figure 2: Measurement of the bits per second

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 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.[13] 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/from Harvard's 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.

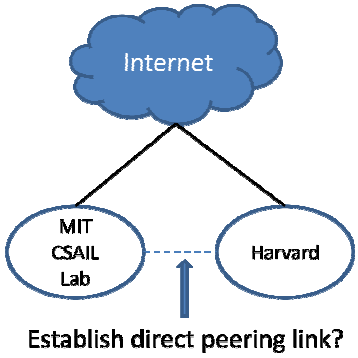

Figure 3: Traffic matrixes derived from Netflow style data is one way of determining where peering links should be established.

Another way of reducing link utilization is to constrain the top contributors to traffic on a link. The figure below is a list of top traffic contributors, again derived from Netflow data. At our lab

[13] Netflow is the common name for this record type but it has been standardized now as Internet Protocol Flow Information eXport (http://en.wikipedia.org/wiki/IPFIX).

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 the network resources. These tools provide a way of objectively identifying "hot spots" and measuring the impact of different experiments, web sites, and uses.

Average rankings for the last 37 topN reports

Top 10 by **bytes in**
built on aggregated topN 5 minute average samples to date

rank	in Address	bits/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 samples)	60.0 M	1.4 M	6.3 k	3.2 k	223.3 m	233.3 m
#2	infinite-state.csail.mit.edu 128.30.24.177 (1 samples)	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 samples)	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 samples)	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

Figure 4: Top contributors to traffic on CSAILs lab network for one period in September 2009.

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 5 below shows a sample of the protocols in use on our lab network.) Particularly as new applications and services that are video-centric become more popular, monitoring their adoption will be key to capacity planning.[14] 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

[14] 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 informs planning to ensure an appropriate quality of service for the applications that are expected.

both within the packets and other predictable signatures such as the pattern of communication (bytes transmitted during the initial connection handshake, etc).[15]

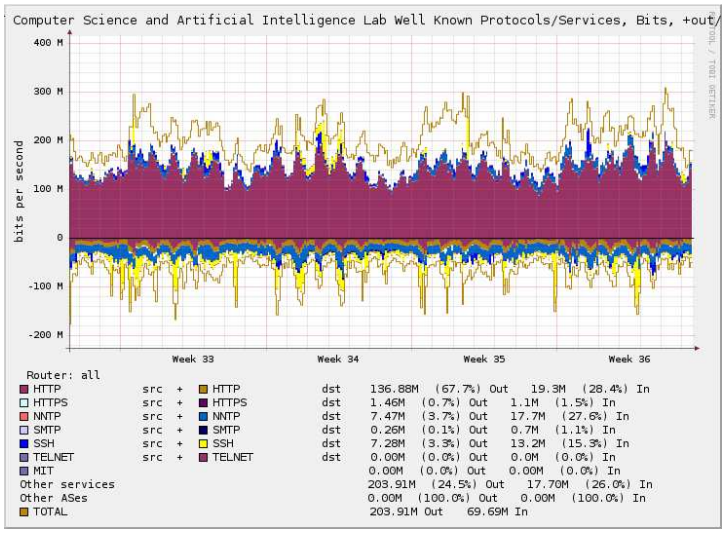

Figure 5: A sample of protocols in use on the MIT CSAIL network. Most of the outbound traffic is http (i.e. web based traffic).

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 via interconnection arrangements of various sorts (e.g., peering or transit). 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.[16] That project represents the most advanced publicly-reported broadband data project undertaken to date -- seven large ISPs, carrying roughly 40% 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.

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

---

[15] 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.

[16] Cho, Kenjiro *et al.* (2008), "Observing Slow Crustal Movement in Residential User Traffic," presentation slides, August 2008, see http://www.caida.org/workshops/wide/0808/slides/residential_user_traffic.pdf; or Cho *et al.* (2006) "The Impact and Implications of the Growth of Residential User-to-User Traffic," paper presented at SIGCOMM 2006, see http://www.sigcomm.org/sigcomm2006/discussion/showpaper.php?paper_id=21.

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 PM and 11 PM. (See Figure 6.) This is important to 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.[17] This has obvious implications for cost sharing and serving pricing.

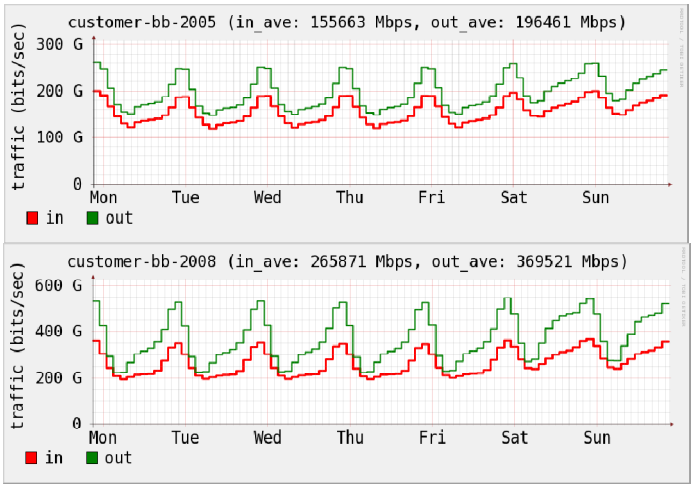

Figure 6: Residential broadband traffic in May 2005 (top) and May 2008 (bottom) as measured in Japan. Source: (Cho 2008).

The MINTS traffic study, run by Andrew Odlyzko at the University of Minnesota, has been monitoring traffic growth levels on networks for a number of years.[18] The traffic data in his study is derived from publically 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 MRTG and RRD graphs (very similar to the previous figures in this paper).[19] 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 available. The MINTS data shows that the aggregate level of traffic continues to grow at double-digit rates, recently averaging around 50-60 percent CAGR

[17] See (Cho 2008).

[18] See http://www.dtc.umn.edu/mints/home.php.

[19] MRTG stands for Multi Router Traffic Grapher and RRD stands for Round Robin Database.

per year.[20] While these growth rates are impressive, they are substantially below the rates widely cited in the trade press over the years.[21]

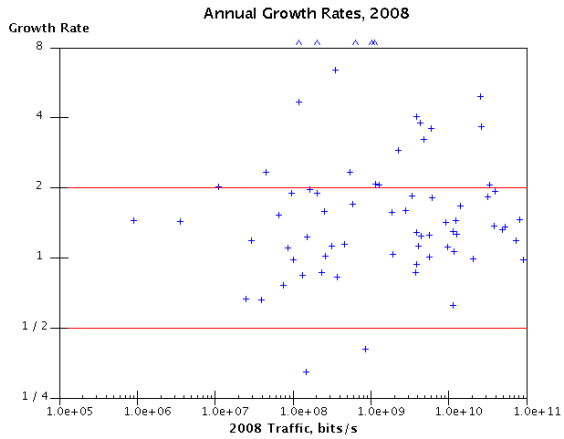

Figure 7: Traffic growth rates from publicly observed sites in the MINTS traffic study.[22]

Cho (2008) 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[23] 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." This is a good metaphor as the increasingly 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 of new applications and services, competition, the value chain of network vendors and suppliers, etc.

The final question, which existing public data sheds some light on, 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

---

[20] See http://www.dtc.umn.edu/mints/home.php.

[21] For example, see "Net traffic doubling every six months," a report from August 2001 (see, http://www.theregister.co.uk/2001/08/17/net_traffic_doubling_every_six/).

[22] See, Minnesota Internet Traffic Studies (MINTS) at http://www.dtc.umn.edu/mints/home.php, for data on traffic growth rates.

[23] See Cho (2008).

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 2 gigabytes of traffic in a month, the top users on a system can easily exceed 100 gigabytes. Figure 8 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.

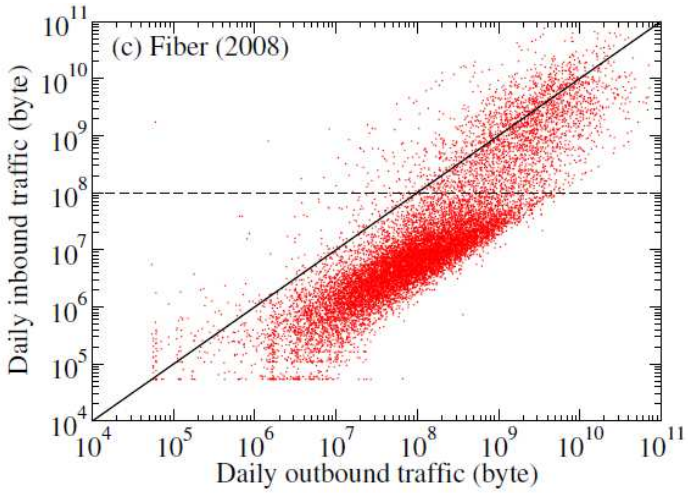

Figure 8: Correlation of daily inbound and outbound traffic volumes per user in one Japanese metropolitan prefecture for a fiber optic network in 2008.[24] Each dot above the dashed line represents users that sent more than 100 megabytes of traffic in a day.

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.[25] 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 (such as usage sensitive charges from an upstream network provider).

[24] See Cho (2008).

[25] 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.

3. 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.[26]

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.[27] By pooling/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 Internet[28] can be developed while still protecting the confidentiality of individual network operators and subscribers.

In particular we see this data as 1) important to establishing traffic trends/growth/characterizations at both the aggregate and subscriber level; 2) vital inputs into a data driven discussion of network management practices; 3) promoting public and industry awareness of the challenges, successes, and opportunities in the broadband marketplace; and 4) assisting in diagnosing and understanding traffic problems and phenomena.

3.1. 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

---

[26] This is not to say there are not a number of researchers addressing this challenge. For example, in addition to the work by Odlysko/MINTS and Cho/Japanese ISPs, see the work of Caida 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 U.S. and abroad.

[27] The wider community includes non-operator participants in the industry value chain (e.g., equipment, application, content providers, and value-added service providers), end customers, third party analysts, and policymakers.

[28] 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.

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 indentify local phenomena, and to discern the drivers and relationships among the sub-components. Data on top line growth alone is not adequate to address questions about the changing mix of applications (e.g. p2p vs. streaming video), differences in platform technologies (e.g., cable modem v. DSL v. wireless), and/or changes over time (in response to changing technology/network architecture and the industry ecosystem). Data to allow the decomposition of aggregate growth are needed for the development of rich future scenarios and to support flexible "what-if" analyses. End-user traffic data is needed to understand "within" and "across" end-user traffic distributions (e.g., how do subscriber usage patterns vary across subscribers and across time?). When linked to cost/revenue data, representative traffic data underlies a fuller understanding of broadband economics.

3.2. 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,[29] significant cable cuts,[30] routing incidents,[31] major media events,[32] and security incidents.[33] 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 end users.

3.3. 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. In his book *Code and Other Laws of Cyberspace*,[34] Lawrence Lessig explored the ways in which code could be an instrument for social control, leading to his dictum that "Code is

---

[29] Internet Captivity and the De-peering Menace http://www.renesys.com/tech/presentations/pdf/nanog-45-Internet-Peering.pdf

[30] 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

[31] The Day the YouTube Died: What happened and what we might do about it http://www.renesys.com/tech/presentations/pdf/nanog43-hijack.pdf

[32] Akami's "Net Usage Index for News" enables users to monitor global news consumption 24 x 7, seeing in real-time the impact of current events on online media consumption. http://www.akamai.com/html/technology/nui/news/index.html

[33] Conficker/Conflicker/Downadup worm as seen from the UCSD Network Telescope http://www.caida.org/research/security/ms08-067/conficker.xml.

[34] Basic Books (July 13, 2000)

law". He might equally have observed (thought less pithily) that network management can be law. The community of network stakeholders therefore needs better ways of understanding and evaluating these policies.

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 provider selected congestion management policies.[35] These policies often change the network resource allocations that would result from the distributed actions of hosts' 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.[36]

3.4. 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.[37] 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/pricing models, *and* better network management.

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.[38] More granular data about subscriber usage patterns would help

---

[35] For example, regulatory authorities have initiated proceedings to examine ISP traffic management practices (e.g., in Canada see http://www.crtc.gc.ca/ENG/archive/2008/pt2008-19.htm, November 2008; and in the U.S. see http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-08-92A1.pdf, January 2008).

[36] For further discussion of why these issues are contentious, see Bauer, Clark, and Lehr (2009).

[37] Broadband Working Group (2005), "Broadband Incentive Problem," a white paper by the MIT Communications Futures Program, September 2005 (available at: http://cfp.mit.edu/publications/CFP_Papers/Incentive_Whitepaper_09-28-05.pdf).

[38] See, for example, Varian, Litan, Elder, and Shutter (2002); Lehr, Osorio, Gillett and Sirbu (2005); Greenstein & McDevitt (2008); or, Dutz, Orzag, and Willig (2009).

improve these studies and offer insight into how best to promote universal adoption and how best to target public broadband funds.[39]

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.[40]

4. 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 non-trivial because forging even temporary agreement on a methodology requires the assent of the technical, legal, and management teams of all participating organizations.[41]

4.1. 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 go back and get the past data thereby leaving holes in the data record.[42] On the point of ambiguous fields, in discussions with network operators, it was interesting to learn that even they do not always fully understand in detail how the measurements of traffic on their network is being done. The operators sometimes need to query their measurement equipment vendors to determine exactly how some of the measured values are calculated and the relevance of particular reported statistics (e.g., the frequency with which fields are updated, or the need to disregard certain statistics because other phenomena such as changes in the measurement design have rendered them no longer meaningful).

---

[39] For example, the American Recovery and Reinvestment Act of 2009 (Pub. L. 111-5, 123 Stat. 115, 2009) has targeted $7.2 billion in public funding for the promotion of broadband.

[40] See http://www.connectivityscorecard.org/images/uploads/media/TheConnectivityReport2009.pdf.

[41] In the Japanese study of Cho (see note 16), the challenges were described as mainly political not technical.

[42] To economize on data storage costs, raw data is summarized in real-time.

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. For instance, fewer measurement probes are need if they are located further up the link aggregation hierarchy, however these will miss intra-node direct communication traffic that occurs 'below' the measurement points in the hierarchy. In the case of analysis boxes (or, DPI) 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"). For instance, traffic classification rules initially did not identify streaming video over the customary TCP port of 80 in its own separate category. Traffic classification techniques differ across equipment vendors so 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 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.

4.2. 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, etc) a subscriber takes, the advertised service characteristics (peak rate, service pricing, etc.) 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.[43]

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

[43] This is information that is not available in the Japanese studies mentioned in note 16.

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 "right" metrics for many of the questions that come up in discussion of traffic data. In a related paper,[44] 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 OECD report that documents the "fastest advertised broadband speeds" per country in 2008. 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/services, the rates that are advertised in Japan and the U.S. are different. For example, the Japanese advertise a maximum peak rate of 160 mb/s that is not achievable in practice.[45] Moreover what is being advertised is the shared capacity. While a subscriber may occasionally be able to burst at very high rates, they are unlikely to sustain such high rates as they compete for the shared resource with other subscribers. Most users appear to understand this and so the advertised peak rate provides, at best, only a rough proxy for expected service quality. Furthermore, we have been told that Japanese subscribers reportedly see broadband connection speeds as status symbols so advertising high peak rates is a valid response to market conditions in Japan. In the United States, the advertized maximum rates for the same technical system are noticeably lower.

[44] See Bauer, Clark, and Lehr (2009).

[45] See Figure 9 below, noting in particular Japan, which offers 160 mb/s broadband. This particular service is offered at 6,000 yen ($60, according to 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 4 channels are bonded together the maximum synchronization speed of the subscribers modem and the CMTS is around 170 mb/s but the maximum usable speeds is 152 mb/s (i.e. less than the advertised speed.).

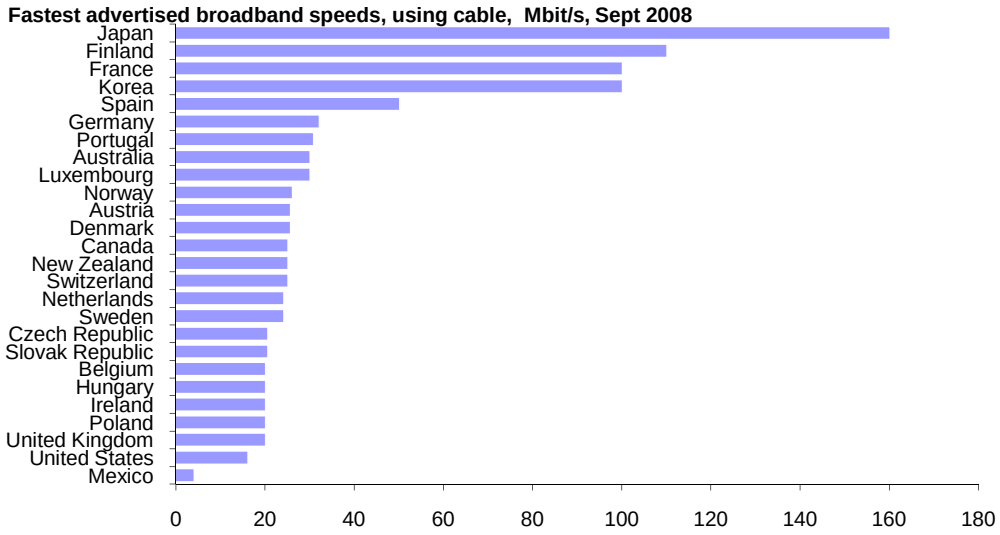

Figure 9: OECD data from 2008 on the fastest advertised broadband speeds.

4.3. 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 individual confidentiality and privacy. Today, there are no clear or universally accepted norms or rules for protecting user data on the Internet. This is an active and important area of on-going research.[46]

Even in the absence of consensus norms or rules,[47] 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 or not).[48] 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

---

[46] See for example, Camp (2001), WIK-Consult *et al.* (2007), or Ohm (2009) for a sampling. Links to current policy debates and further research are available at: Electronic Privacy Information Center (http://epic.org/), privacy.org (http://privacy.org/), Electronic Frontier Foundation (http://www.eff.org/issues/privacy).

[47] ISPs do face regulatory rules on the disclosure of subscriber data, but these rules vary by context and are not comprehensive.

[48] In today's Internet blogosphere rumors of bad behavior can be damaging.

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)[49] 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,[50] 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).

4.4. 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.[51] 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. While the data *could* be compelled, using strong regulation to collect/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 than if its collection is only accomplished via regulatory mandate.[52]

---

[49] The IRB is a review board established by the University with representation from across the community.

[50] These included the Milgram Obedience to Authority experiments of 1988 and the Public Health Service Syphillis Study (1932-1972), more commonly referred to as the "Tuskegee Syphillis" experiments.

[51] See, for example, Tirole (1988).

[52] Government mandated data collection of detailed data comes with strong non-disclosure obligations. For example, the Annual Survey of Manufacturers (ASM) collected by the Census Bureau limits disclosure of data that would allow identification of individual providers. To date, the FCC has severely limited third-party access to the zip code level data it has collected on broadband, thus limiting its usefulness for policy formulation and research (lots of cites to this possible if we need).

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.

5. Collaborative data collection

Over the past year, we have been working with a group of broadband providers to collect "micro" traffic data sets that we plan to pool in order to provide a richer characterization of broadband traffic than has heretofore been publicly available.[53] We are aware of a number of complementary initiatives also underway.[54] Our project, the MIT Internet Traffic Analysis Study (MITAS), is collecting a pooled data set from a number of ISPs serving a cross-section of geographically dispersed markets using a variety of network architectures in the United States and abroad.

The data we are collecting falls into a variety of different categories. In almost all cases this is data that network providers are collecting internally already. The data we are collecting includes:

- Per subscriber usage data (up and down byte counts over different measurement periods ranging from 15 minutes, 1 hour, to 1 month) from sources such as IPDR data records[55]
- Link utilization records that represent the aggregate data flow for subscribers (up and down byte counts over different measurement periods again ranging from 15 minutes, 1 hours, to 1 month) collected from sources such as SNMP byte counters
- Percentages of traffic in different classifications such as http, email, video, peer-to-peer etc as determined by the rules of 3rd party traffic analysis boxes.
- Historical utilization data at both for both subscriber level and aggregate data flows
- Reports detailing interesting traffic management incidents and challenges
- Results of individual experiments that for instance explore the correlation between link utilization and packet loss, latency, and jitter.

To give a flavor of some of the research questions we believe this data may help to address, consider the following partial list of questions we are considering targeting in our initial research efforts with this data:

---

[53] See http://mitas.csail.mit.edu for further details.

[54] Some of the important initiatives in this area include MINTS (http://www.dtc.umn.edu/mints/); CAIDA (http://www.caida.org/research/traffic-analysis/), and M-Lab (http://www.measurementlab.net/).

[55] See http://www.tmforum.org/ipdr for details.

- What are the traffic growth rates in different parts of the Internet?
- What does the peak to average ratio look like in different markets and regions of the world?
- Do the top users actually consume more bandwidth at peak times than the average users?
- What share of subscriber traffic terminates on-net?
- How do individual subscribers' usage levels change over time (during a subscriber's lifecycle, seasonally, and with general industry trends)?
- What is the impact of external events on overall traffic (e.g., special events such as the Olympics, extreme weather such as heat waves, or the introduction of new services/applications/devices such as YouTube or the iPhone)?
- What is the relationship, if any, between global security threats, botnets, and malicious traffic, and bandwidth consumption patterns?
- What are useful summary statistics and data representations to characterize traffic aggregates? Per-subscriber flows?

6. 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, etc.),[56] and the applications enabled by broadband become more widely relied upon (e.g., interactive rich multimedia applications),[57] questions about *how* broadband is being used will be more interesting than whether it is being used. To properly address such questions, we will need much better insight into Internet traffic (its growth, statistical characterization, drivers, etc.) over both short (operational) and longer (investment) time frames.

We have argued in this paper that traffic data will be central to monitoring and resolving the inevitable tussles of this next stage or development. Given that such data is not publically visible, the cooperation of network operators is essential. We are optimistic that the challenges of sharing

---

[56] With first generation broadband, knowledge of the basic technology (e.g., DSL or cable modem) provided a rough but reasonable gauge to understand what the peak data rates might be (upstream/downstream) and what applications could be supported via those services. As we move to a world with FTTH and 4G wireless, there will be wider dispersion in available peak rate capabilities in marketed services that will render reliance on peak rates alone a much less reliable metric for characterizing service differences. Traffic data becomes more important to develop a nuanced understanding of relevant architectural differences.

[57] As noted in the preceding footnote, there are many ways to provision broadband capacity for different applications and what works best for one application may not for another. Once again, a richer understanding of traffic data is essential to adequately mapping services to available physical and network infrastructures.

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,[58] NANOG[59] (and its equivalents in other regions), and other forums. A similar cooperative capacity can be developed which produces data about the 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".[60] 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.

7. References

Bauer, S., D. Clark, and W. Lehr (2009), ""The Evolution of Internet Congestion," paper prepared for 37th Research Conference on Communication, Information and Internet Policy (www.tprcweb.com), Arlington, VA, September 2009.

Briscoe, B. (2007), "Flow Rate Fairness: Dismantling a Religion." *SIGCOMM Comput. Commun. Rev.* 37, 2 (Mar. 2007), 63-74.

Camp, J. (2001), *Trust and Risk in Internet Commerce*, MIT Press: Cambridge, MA 2001.

Cho, K., Fukuda, K., Esaki, H., and Kato, A. (2008),"Observing slow crustal movement in residential user traffic," In Proceedings of the 2008 ACM CoNEXT Conference (Madrid, Spain, December 09 - 12, 2008). CONEXT '08. ACM, New York, NY, 1-12.

CSTB (2001), *Looking over the Fence at Networks: A Neighbor's View of Networking Research* (Natl Academy Pr, 2001).

Dutz, Mark, Jonathan Orszag, and Robert Willig (2009), "The Substantial Consumer Benefits of Broadband Connectivity for U.S. Households," Compass Lexecon, A Study Commissioned by the Internet Innovation Institute, July 2009.

[58] Internet Engineering Task Force (see http://www.ietf.org/)

[59] The North American Operators Group (see http://www.merit.edu/nanog).

[60] See CSTB, 2001.

Faratin, P., D. Clark, P. Gilmore, A. Berger, and W. Lehr (2007), "Complexity of Internet Interconnections: Technology, Incentives and Implications for Policy," paper prepared for 35th Annual Telecommunications Policy Research Conference, George Mason University, September 2007.

Floyd, S. and M. Allman (2008), "RFC 5290: Comments on the Usefulness of Simple Best-Effort Traffic," Network Working Group, Internet Engineering Task Force, July 2008, available at: http://www.faqs.org/rfcs/rfc5290.html.

Fukuda, K., Cho, K., and Esaki, H. (2005),"The impact of residential broadband traffic on Japanese ISP backbones," Sigcomm Comput. Commun. Rev. 35, 1 (Jan. 2005), 15-22.

Greenstein, Shane and Ryan C. McDevitt (2009), "The Broadband Bonus: Accounting for Broadband Internet's Impact on U.S. GDP," White paper, Technology Policy Institute, Washington, D.C., January 2009 (available at: http://www.techpolicyinstitute.org/files/greenstein-broadband bonus1.pdf)

Jacobson, V. and M. J. Karels (1988), "Congestion avoidance and control," ACM Computer Communication Review; Proceedings of the Sigcomm '88 Symposium in Stanford, CA, August, 1988, vol. 18, 4, pp. 314-329, 1988.

Lehr, William, Carlos Osorio, Sharon Gillett, and Marvin A. Sirbu (2005) "Measuring Broadband's Economic Impact," paper prepared for Telecommunications Policy Research Conference, Arlington, VA, September 2005.

Lehr, William, Jon Peha, and Simon Wilkie (eds) (2007), *Special Section on Network Neutrality: International Journal of Communication* (volume 1, 2007), August 2007 (http://ijoc.org/ojs/index.php/ijoc/issue/view/1).

Leiner, B.M., V.G. Cerf, D.D. Clark, R.E. Kahn, L. Kleinrock, D.C. Lynch, J. Postel, L.G. Roberts, and S. Wolff (1997), "A Brief History of the Internet," *Communications of the ACM*, vol. 40, 1997.

Licklider (1963), "Topics for Discussion at the Forthcoming Meeting, Memorandum For: Members and Affiliates of the Intergalactic Computer Network". Washington, D.C., Advanced Research Projects Agency, 23 April 1963 (via KurzweilAI.net, retrieved 2009-06-15).

Lyon, M. (2004), *Where Wizards Stay Up Late*, Simon & Schuster: New York, 2004.

Mathis, M. (2009), "Rethinking TCP Friendly", March 2009 at http://staff.psc.edu/mathis/papers/TSVAREA73.pdf (retrieved March 30th 2009).

Ohm, Paul (2009), "The Rise and Fall of Invasive ISP Surveillance," *University of Illinois Law Review*, August 2009.

Peterson, L. L. and B.S. Davie (2007), *Computer Networks: A Systems Approach*, Fourth Edition, Morgan Kaufmann: New York, 2007.

Stallings, W. (2008), *Operating Systems: Internals and Design Principles*, 6th Edition, Prentice Hall: New Jersey, 2008.

Subramanian, M, *Network Management: Principles and Practice*, Addison Wesley: New Jersey, 1999.

Tirole, J. (1988), *The Theory of Industrial Organization*, The MIT Press: Cambridge, MA, 1988.

Varian, Hal, Robert Litan, Andrew Elder, and Jay Shutter (2002), "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.

Wik-Consult/Rand Europe/CLIP/CRID/GLOCOM (2007), *Comparisons of Privacy and Trust Policies in the Area of Electronic Communications, Final Report*, report prepared for the European Commission, July 2007 (available at: http://ec.europa.eu/information_society/policy/ecomm/doc/library/ext_studies/privacy_trust_policies/final_report_20_07_07_pdf.pdf).