Measuring Broadband:
Improving Communications Policymaking through Better Data Collection

Kenneth Flamm, Amy Friedlander, John Horrigan, and William Lehr

A report of workshop co-sponsored by
Pew Internet & American Life Project
University of Texas at Austin, with support from the National Science Foundation
The Massachusetts Institute of Technology

Convened on June 28, 2006
The Pew Research Center
Washington, DC USA

November 2007

Table of Contents

This essay is based upon a day-long workshop organized by Kenneth Flamm of the University of Texas at Austin, John B. Horrigan of the Pew Internet & American Life Project, William Lehr of the Massachusetts Institute of Technology, and Sharon Gillett of MIT (at the time of the workshop). The text was written by Amy Friedlander, in collaboration with Kenneth Flamm, John Horrigan, and William Lehr.

This document should be cited as Kenneth Flamm, Friedlander, Amy, Horrigan, John B., Lehr, William. *Measuring Broadband: Improving Communications Policymaking through Better Data Collection*. (Washington, D.C.: Pew Internet & American Life Project, 2007). The views and opinions expressed herein are those of the authors. They do not necessarily reflect the views and opinions of the Pew Research Center, the Pew Internet & American Life Project, the Massachusetts Institute of Technology, the University of Texas at Austin, the National Science Foundation or the U.S. government.

Executive Summary

Questions of vital importance to understanding the information society are difficult to address because of poor data.

- Do places with more widely available or higher quality information infrastructure perform better economically than those without?
- Does new investment in broadband connections have economic or social payoffs for communities?
- Are civic institutions healthier in broadband rich areas – or not?
- What portion of the lag in home broadband adoption in rural America is attributable to lack of available infrastructure?
- Do those with "second generation broadband," such as those with fiber to the home, behave differently online than those with "first generation" broadband such as DSL or cable modems?
- How should we account for the impact of advanced information networks in measures of productivity and other economic activity?

Imperfect or absent data are are rarely mentioned in policy discussions. Yet the communications policy debate in the United States today is inseparable from debates about the data used to make claims about policy propositions. President Bush articulated in 2004 a goal to have universal and affordable broadband available in the United States by 2007. The way data are collected by government agencies cannot answer questions about whether that goal has been met or not. International organizations – using the imperfect data – report that the United States' ranking in per capita broadband adoption is lower today than it was a few years ago. This paper argues that the country cannot properly gauge its own progress or know how dire America's international standing is without good data about broadband adoption, deployment, price, and quality.

In June 2006, researchers from the Pew Internet & American Life Project, the University of Texas at Austin, and the Massachusetts Institute of Technology convened a workshop of like-minded specialists from government, academia, and industry to discuss challenges involving the state of data collection about the deployment and use of communications infrastructure. The workshop's wide-ranging discussions yielded the following recommendations on the principles that should guide efforts to improve data collection on the deployment and use of communications infrastructure.

- **Collection of data should be at a sufficiently fine-grained level to permit regional analysis of the impacts of communication technology.** Nearly all publicly available data on adoption of communications goods and services are gathered at the national level. More granular data – about adoption patterns among individuals and businesses at the local or regional level – would permit more rigorous analysis of the impacts of information and communications technology on economies and communities. Such data should capture the price users pay for service. With such data, policy makers and researchers would be better able to determine the expected payoffs from encouraging broadband deployment and adoption.
- **The United States should be able to produce a map showing the availability of infrastructure in the country.** The current methods of tracking the availability of high-speed infrastructure relies on providers reporting by 5 digit zip code where they offer service. A provider with one customer in a zip code can report that it provides service in the zip code, which may misleadingly suggest that the entire zip code can get service.

Understanding where infrastructure is available is critical for understanding what kind of choices consumers have for service. Workshop participants argued thaqt efforts to improve mapping of infrastructure must be accompanied by the Federal Communications Commission updating the definition of broadband to reflect advances in the nation's information infrastructure. The current decade-old definition of 200 kilobits per second for broadband is widely viewed as outdated.

- **Academic researchers, non-profit organizations, the government, and the private sector must work collaboratively to gather data that permits assessment of quality of service and the user experience**: The type of internet experience the end-user has at the desktop depends on "last mile" infrastructure availability, users' awareness of it, and capacity to securely and skillfully take advantage of online connections. Assessing quality of service depends, in part, on computer scientists measuring online data traffic. These online metrics have become more difficult to acquire since the internet backbone was privatized, yet they remain important to properly maintaining an increasingly vital part of the nation's critical infrastructure. Understanding the user experience also requires social science research into the community and cultural contexts of technology adoption and use.

The 2006 workshop was motivated by the desire of researchers to have better data with which to study the social, economic, and policy consequences of dissemination of information and communications technologies. At the same time, workshop participants recognized the sensitivity endemic in collecting data on commercial activity. Companies understandably do not want to make available to the public data that might reveal proprietary or strategically important information.

Yet the rewards from confronting those challenges and improving data collection are great. Policymakers would have a better understanding of the social and economic consequences of investments in communication infrastructure they may have under consideration. Economists in the public and private sector could better understand how new information and communication technology affect productivity. And with a clearer understanding of user behavior, planners in government agencies at the state, local, and federal level could more effectively design electronic service delivery applications for citizens.

Since the workshop, both the U.S. House of Representative and the Senate have held hearings on bills designed to improve data collection on broadband infrastructure. These bills represent valuable first steps in addressing this issue. Continued dialogue between researchers and policymakers is needed to develop data collection practices in the United States that will allow for informed deliberations on communications policy.

Introduction

The workshop on broadband metrics that is discussed here was held in June 2006, and in light of recent events, was either prescient or instrumental in helping to mobilize wider support for improving the state of our collective public knowledge of broadband networks. In May 2007, Commerce Committee Chairman Daniel Inouye (D-Hawaii), with a number of co-sponsors, introduced the Broadband Data Improvement Act (S.1492) that is designed to improve federal and state broadband data collection. According to Senator Inouye, "The first step in an improved broadband policy is ensuring that we have better data on which to build our efforts."[1] The Senate Commerce Committee reported the bill out of Committee for consideration by the full Senate in July. In October 2007, the House Subcommittee on Telecommunications and the Internet reported out the Broadband Census of America Act to improve data collection on high-speed internet availability.

There is no disagreement among technology-policy makers that broadband is a basic infrastructure that is critical to the health of our economy and social life. Infrastructure and services continue to evolve, with the continued growth in penetration of first generation DSL and cable modem services, the expanded availability of mobile 3G broadband and nomadic WiFi broadband, as well as fiber-to-the-home services. This creates the need for better data to track the progress and impacts of broadband service nationwide. In what follows, we describe the efforts of leading broadband researchers to provide a snapshot of the broadband data debate as it looked in June 2006.

Only ten years ago, it made sense to ask, who had internet access and who did not? Now we ask, how fast is your connection? And how fast is fast? The Federal Communications Commission (FCC) currently defines high speed service as greater than or equal to 200 Kilobits per second (Kbps) in one direction, a decision announced in agency's first report on broadband deployment, required by the Telecommunications Act of 1996. The 200Kbps metric was selected for a number of reasons, including the desire to pick a data rate that would reflect a significant improvement in dial-up connections operating at 50Kbps and would exclude ISDN connections at 128Kbps. ISDN, in 1996, was generally available and marketed as an advanced service, but it was never widely adopted, in part because of high usage-sensitive pricing. The 200 Kbps metric would include most other emerging services commonly seen as high-speed internet access at the time.[2]

Today, broadband services offering peak download rates measured in several Megabits per second (Mbps) are common, and new offerings based on fiber-to-the-home are being deployed that are capable of supporting 10's of Mbps data connectivity. Ten years ago, wireless internet connections were exotic; now they are an amenity at corner coffee shops, hotels, and terminals at

---

[1] See "Inouye introduces Broadband Data Improvement Act," Press Release, United States Senate Committee on Commerce, Science, and Transportation, May 24, 2007 (available at: http://commerce.senate.gov/public/index.cfm?FuseAction=PressReleases.Detail&PressRelease_id=248822&Month=5&Year=2007).

[2] U.S. Federal Communications Commission Inquiry Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996, CC Docket No. 98-146, January 28, 1999, p. 20, http://www.fcc.gov/broadband/706.html
U.S. Federal Communications Commission.

major airports. There have also been significant in end-user equipment (e.g., routers supporting data rates in excess of 50Mbps are available for less than $50). In light of these developments, the FCC has begun to reconsider its data collection policies.

In addition to definitional worries about broadband, various stakeholders increasingly ask about broadband deployment: Why is broadband not adopted by some residents when it is available? How is broadband used by subscribers? Policymakers and community officials also inquire about the state of competition in broadband markets and provision of broadband from alternative service providers, including mobile broadband over 3G networks, fixed wireless broadband, broadband-via-power lines, and fiber-to-the-home deployments. There is also significant variation in broadband adoption in population sub-segments. Although 73% of American adults were internet users at the time of the workshop, there are still demographic groups and locales (mostly rural) where service options are non-existent or limited, and where usage rates are significantly below the national average. And despite recent rapid advances, especially in Africa and Latin America, there remains a global digital divide in both access and quality of service.

Origins of the workshop

At the Telecommunications Policy Research Conference in the fall of 2005, a group of experienced investigators who were probing the deployment of broadband service from different perspectives discovered they shared a frustration: They were finding that the data on which their respective analyses relied were flawed, limited, and in some instances inappropriate. This constrained the kinds of questions they sought to answer and biased findings that, in turn, could affect public policy decisions. The group of researchers felt that the public bureaucracies that collect data and generate statistics, which are widely used, are inherently conservative and slow to employ new methodologies that might provoke criticism.

The outcome of this conference encounter was a one-day invitation-only meeting in Washington the following June, sponsored jointly by the Pew Internet & American Life Project; the University of Texas at Austin, with support from the National Science Foundation; and the Massachusetts Institute of Technology. Sixty-five people participated as speakers, panelists and members of the audience in a program of prepared sessions and open panel discussions, allowing for lively exchanges.[3] This essay describes issues raised by the speakers and participants and recommendations for going forward. Its focus is measurement, data, and the big questions that are important to formulating public policy and drive current research on broadband. The speakers looked primarily at broadband in the U.S., although several presentations provided international comparisons.

The big and unanswered questions

In laying out the important data collection questions, workshop participants touched on five themes, to be discussed in detail in this section of the essay:

- **Productivity**: Why are accurate measures of broadband and other information and communication technologies (ICTs) important to measuring the economic productivity?

- **Public policy and government intervention**: If government chooses to intervene in the communications market place (e.g., to fill gaps in infrastructure provision), is the

[3] The list of participants and the agenda are included as an appendix to this document.

necessary data available to help government officials make these decisions and assess their impacts?

- **Measuring penetration rates**: As technology continues to evolve rapidly, how should the government and other entities address the challenges in accurately measuring the technologies people have and how they use them?

- **The internet and geography**: If the internet has the potential to overcome geographic barriers, what data are needed to assess claims about the internet's impact on urban or rural development?

- **Culture and users' environment**: What research methods, data, and information-gathering strategies are needed to understand user behavior?

Productivity

We see computers everywhere but in the productivity statistics, Robert Solow famously wrote in a book review in the New York Times in 1987.[4] And with a masterful stroke, he coined the term "the productivity paradox." Productivity is the measure of a technology's contribution to the economy and is essential to understanding and analyzing the determinants of economic growth. Twenty years later, much has been learned about the impact of information technologies on productivity.[5] However, as Jack Triplett explained in his opening remarks at the workshop, the advent of broadband affects two key measures: labor productivity and the more sophisticated measure called multi-factor productivity. Both measures have increased since 1995, especially labor productivity, and most of the growth has been concentrated in the service sector. Since, in economic statistics, broadband is classified in the service sector, flawed data collection methods that do not capture the technologies' effects could significantly impact the productivity measures for the entire U.S. economy. Inappropriate representation of broadband distorts both labor and multi-factor productivity measures, and not in the same way (sees Box 1 for further detail).

Getting productivity wrong, as Shane Greenstein said, can affect a number of economic policy decisions, such as interest rates set by the Federal Reserve. Overstating productivity may be equally problematic if it results in excess investment in ICT that might be better directed toward other resources in the economy. Greenstein and his co-author Chris Forman of Carnegie Mellon University agreed that the contribution of broadband to productivity is an important question. But they noted it remains an open question as to the size of the contribution of ICTs to the U.S. economy. As Flamm asked in his introductory remarks, "Is it [broadband] big enough to merit separate measurement, and if not, the obvious question is when?"

In discussing ICTs, broadband, and productivity, workshop participants noted that distinctions between accounting for broadband and, say, personal computers. Personal computers can be counted, and the purchase is a single transaction, although it may trigger other purchases in the form of software, printers, and other hardware devices. In a business, the equipment becomes part

[4] Robert Solow, We'd better watch out. New York Times Book Review: 36, July 12, 1987; as quoted in Jack E. Triplett, The Mismeasurement hypothesis and the productivity slowdown, pages 19-46, in Productivity, inequality, and the digital economy: a transatlantic perspective, edited by Nathalie Greenan, Yannick L'Horty, and Jacques Mairesse (Cambridge, MA: The MIT Press, 2002).

[5] Jorgensen, Dale W. and Kevin J. Stiroh, "Raising the Speed Limit: U.S. Economic Growth in the Information Age," *Brookings Papers on Economic Activity*, 2000 (1), pp. 125-211.

of the inventory of assets, subject to depreciation. Broadband, on the other hand, is a service made available by providers, who must first make significant infrastructure investments before the first consumer can subscribe. For the consumer, broadband is an ongoing cost, like other utilities, not a one-time investment.[6] The decision to acquire broadband is mediated by both availability and cost.

Box 1: Measuring productivity

Labor productivity (LP) is the ratio of output to labor; multi-factor productivity (MFP) is the ratio of output to a weighted average of measures of the aggregated inputs of capital, labor, energy, materials, and services (which includes broadband services).

$$LP = \frac{output}{labor}$$

$$MFP = \frac{output}{weighted\_average(capital, labor, energy, materials, services)}$$

Triplett reported that results from 1995 showed that most productivity improvements were concentrated within services industries, rather than within the goods producing industries. Industries that produce broadband services are in the services sector, where most of the consumption of broadband services also occurs. Based on data from the Bureau of Economic Analysis (BEA), labor productivity in the communications services industry increased 8.4% per year after 1995 and multi-factor productivity grew at the much slower 1.4% per year. However, it is unknown whether broadband data was included in the BEA account. If broadband were not measured appropriately in the output measure, the output measure's impact is too low and labor productivity is too small. When analyzing industries that use broadband, the concern is whether broadband services have been appropriately included in the services (S) term when calculating MFP. In these cases, labor productivity (the ratio of output to labor) may be accurate, but MFP would overestimate the role of broadband in the denominator were understated.

The local nature broadband service presents additional challenges. Broadband is, after all, mainly a wireline service provided at a particular geographic location. Consequently, it is reasonable to expect that part of its direct impact would be local (as well as any spillover benefits that may accrue over larger areas). Greenstein and Forman said that the best way to measure direct economic impacts of broadband is to focus on use by business establishments. Household use may have indirect effects on the economy but measuring it is more difficult. They posed three clusters of largely empirical questions that prompt examination of broadband use by industries and firms and help determine the economic impacts:

[6] This distinction is drawn by Schement and Forbes in the context of the comparative history of telephony, radio, and television; they extrapolate from those examples to the internet. See Jorge Reina Schement and Scott C. Forbes. Identifying Temporary and Permanent Gaps in Universal Service, The Information Society 16 (200): 117-26.

- What industries make the greatest use of broadband? How has this impacted their productivity? Which types of firms are most strongly impacted when broadband becomes available? When broadband prices decline or quality improves?

- Which areas of the country have been most affected by broadband in the last decade? Which areas would benefit most from increasing availability, declining prices and quality improvements in broadband? Which areas may have suffered from broadband's diffusion (e.g., might a region whose comparative advantage depended on communications infrastructure suffer from the widespread deployment of high-speed networks elsewhere?)?

- What technologies generally complement broadband? What clusters of technologies are needed to realize broadband's benefits? What investments follow broadband? How does broadband impact spatial organization of productive activities? Firm organization? How has business use of broadband shifted in response to concerns over information security? Has this varied by industry?

Some studies that address these questions are already underway. Since the 1990s, there has been active interest in collecting demographic information on computer use at the household level, led by the work of the Census Bureau. Furthermore, the increased use of remote access, wireless connectivity, and always-on connections are beginning to blur the distinction between life at work and life at home. Avi Goldfarb of the Joseph L. Rotmann School of Management at the University of Toronto outlined questions on individual, household, and commercial uses. Goldfarb addressed questions such as, "Where are American goods finding markets when there are no transport costs?" and "What goods are Americans buying?" He speculated that clickstream data, captured at the desktop (or keyboard), could prove helpful in understanding online behavior on a more intimate and granular scale than has heretofore been feasible.

Sharon Gillett, at the time with the Massachusetts Institute of Technology and now chairman of the Massachusetts Department of Telecommunications and Cable, together with her colleagues William Lehr, Carlos Osorio, and Marvin Sirbu, described research that was focused on measuring the impact of broadband on local economies. Using a zip-code-level panel data set based on broadband availability from the FCC, they examined the impact of broadband availability on the growth in employment, establishments, and industry composition. Their research finds that broadband contributes significantly to increasing the growth rates of all three. However, whether the higher growth rates represent a one-time or permanent improvement is unclear.

Problems with using zip code data to understand determinants of broadband deployment were addressed by Flamm, Tony Grubesic, a geographer now at Indiana University, and James Prieger of the University of California, Davis. They all concurred that the zip code data have a number of significant problems for analysis. These range from inconsistent zip code mappings, changes in zip codes over time, and difficulties in matching across various data sets (e.g., from the FCC data to the Census demographic data organized according to ZCTAs, the Census Bureau's attempt to map its data into zip codes). The data collected by the FCC are critical and widely used. All of the speakers described problems with the FCC's definition of broadband (greater than or equal to 200Kbps in one direction) and the definition of zip codes. Since the FCC is the principle source for national information on the geographic availability of high speed lines at relatively granular spatial units, difficulties in interpreting this data pose a serious challenge for research. Problems with this dataset are discussed in some detail in later sections of this essay.

Public policy and government intervention

Expanding availability of broadband infrastructure requires investment in both hardware and software. The case for such investments, whether undertaken by business or government, requires reliable data on the costs of deploying infrastructure and on expected consumer demand. Moreover, consumer demand changes over time as complementary goods make the service more attractive and as the population of broadband users matures from early adopters to mass market consumers. Thus, analysis of broadband markets requires both knowledge of the range of choices available in the market place as well as what consumers are doing with the services.

Rahul Tongia, of Carnegie Mellon University, who has studied infrastructure in emerging economies with a focus on technology and technology choices, emphasized this distinction between usage, which is typically measured by penetration rates, and access, or availability of the service. In the U.S., he said, dial-up access is ubiquitous because of the universal availability of telephone service, something not realized in much of the world. Furthermore, in contrast to the U.S., where dial-up and broadband access is available for a fixed monthly fee, in many other parts of the world, internet access is subject to time-metered usage tariffs. For example, in India, DSL costs just over $2 a month but usage is capped. Such policies mean that even where services are available, there may be barriers to usage that will limit demand for infrastructure and services.

Differences in the availability of broadband infrastructure, the regulatory and industry environment, and local demographic characteristics all contribute to significant cross-national differences in broadband penetration. Prieger observed that, according to OECD data, the U.S. is not currently in the top ten nations in terms of broadband per capita penetration. His work explores what might be done to address this situation and whether well-crafted policy will have an impact. Gregory Rosston of Stanford University has worked extensively on narrowband, which has lessons for broadband, and has found that "costs vary greatly across a state depending on cities and subsidy programs." More recently, Flamm has found that programs at the state level affect broadband usage.[7]

To encourage universal availability of broadband, U.S. policymakers launched in 1997 the Schools and Libraries Program of the Universal Service Fund, commonly known as "E-Rate." The E-Rate program is administered by the FCC and provides discounts through a competitive bidding process to assist most U.S. public schools and libraries to obtain affordable telecommunications and internet access.[8] Early assessments, including one by the non-profit Benton Foundation, considered the program a success. But subsequent studies by the Congressional Research Service (2004) and the General Accounting Office (2005) raised questions about its management and utility.[9] Anindya Chaudhuri, a colleague of Flamm's at the LBJ School, described some of the issues associated with measuring the effectiveness of the now-$8 billion E-Rate program using the data from the Current Population Survey (Computer and

[7] Kenneth Flamm, Diagnosing the Disconnect: Where and Why is Broadband Access Unavailable in the U.S.? August 2006.

[8] Universal Service Administrative Company. About the Schools and Libraries Program: Overview of the Program, Last modified November 3, 2006; viewed March 27, 2007. http://www.universalservice.org/sl/about/overview-program.aspx

[9] Congressional Research Service. The E-Rate Program: Universal Service Fund Telecommunications Discounts for Schools (RL32018) by Charmaine Jackson. (2004); General Accounting Office. Greater Involvement Needed by the FCC in the management and Oversight of the E-Rate Program. Dated February 2005. GAO-05-151; Benton Foundation, The E-Rate in America: A Tale of Four Cities (Washington, DC, 2000).

Internet Use Supplement), the National Center for Educational Statistics, the decennial U.S. Census and the Universal Service Administrative Company. A critical unit is the school district. Unfortunately, the school district is not identified consistently across the key datasets so, he concluded, "There is no way to track the funds."

In discussing Chaudhuri's findings, Scott Wallsten of the Progress and Freedom Foundation argued that drawing conclusions about the effectiveness of broadband policies is difficult when the data underlying the analysis are flawed or inconsistent. Wallsten also discussed public policy and broadband deployment in a comparative U.S./international framework and identified four problems (which others echoed):

1. Inconsistent definitions: The FCC definition of broadband as greater than 200Kbps makes it hard to compare broadband adoption across countries.

2. Inadequate competition measures: The focus on broadband is too narrow as other things compete with broadband. For example, many Americans still use dial up. Dial up prices are falling and improvements in dial up are occurring.

3. Poor indicators of speeds: Advertised speeds are not necessarily the same as delivered speeds.

4. Poor measures of prices: There is a lack of information regarding how much people are willing to pay for higher speeds.

Measuring penetration rates

Frequent and reliable measures of broadband adoption at the national, regional, and even local level are important for a number of reasons. Economists seek to understand why broadband services are more available in some locales than others and whether this affects economic performance. Policymakers worry about the implications of "connectedness" on civil and political engagement. Traditionally, such research relies heavily on the public data assembled by the statistical agencies at the federal and, to a lesser extent, state level. This research also uses privately funded, independent surveys. As Greenstein, Forman, and Goldfarb noted, this research typically does not cover business uses, where there is substantial direct economic impact, but tends to examine use at home. This bias toward household use may be appropriate for mature markets, but may be less useful for understanding key adopter communities, since many people learned to use the technology at work or school. In general, the large, national scale surveys have provided good baseline demographic information, but they generally do not lend themselves to more sophisticated analysis of niche markets, like the early adopter communities.

The National Telecommunications and Information Administration (NTIA) has done a series of studies of computer and internet use at the household level, using the monthly Current Population Survey conducted by the Census Bureau for the Bureau of Labor Statistics. However, this program was discontinued in 2004 with the publication of the 2003 data. The Pew Internet & American Life Project, as well as the work done by the Center for the Digital Future at the Annenberg School at the University of Southern California, have continued to examine this topic, as have others, looking beyond basic demographic studies more generally to online behavior.

At the workshop, Horrigan summarized some of the extensive work that the Pew Internet Project has done on broadband penetration and discussed some of the limitations of its survey

methodology.[10] The project relies on a carefully structured, random-digit telephone survey design. Such surveys are appropriate for capturing a number of measures about users' broadband experience, but less so for other important measures, such as network speed available to a user. Horrigan cited a Pew Internet survey that found that 81% of home high speed users could not identify how fast their home connection was.

The internet and geography

One of the initial promises of the internet was the ability to overcome geographic barriers. Productive activities could be distributed remotely and users could communicate via the internet. Today residential users and small business owners may now access remote storage and backup services over the internet, while companies of all sizes may avail themselves of outsourcing services impossible to support in a pre-broadband era. The physical location of the storage facilities is less important and may be optimized to take advantage of the relative trade-offs of falling bandwidth costs versus local real estate and power costs for maintaining the off-site storage facilities. Technical issues such as the choice of storage media and format migration may be outsourced, while issues such as how to protect individual privacy and confidentiality may become more important.

In spite of the internet's promise of reducing the barriers of geographic distance, location continues to matter when economic and public policies are considered. Cross-national differences in data privacy rules or intellectual property protection may influence where data is stored and how it is routed when transmitted. Additionally, scale continues to matter. At the national level, population is not uniformly distributed, and markets for networked services have historically been clustered. In the U.S., telegraph lines and railroads in the 1840s and 1850s connected population centers. Initial deployment of telephony and electrical systems in the 1880s and 1890s followed similar contours. Decisions regarding the placement of infrastructure reflected trade-offs between technological constraints, notably signal attenuation, and profitability. In the early 20th century, constructing electrical power lines cost $2,000 per mile and with an average of 2-to-4 hook-ups per mile, extending lines to rural areas was considered infeasible.[11] Moreover, with time, legacy systems and policy decisions affect availability, as Rosston's work on narrowband suggests.

Local land use policies also play a role in the cost of infrastructure installation. Tongia pointed out that the availability of transit is a big determinant of pricing in rural areas. Issues such as the rights to attach new infrastructure to telephone or electric poles, control over conduit, and rights of way may vary by locale and may pose a significant impediment to the deployment of competitive infrastructure.

Geospatial studies of availability of and access to broadband are another research area in which data problems limit opportunities for important analysis. Indeed, use of geographic information systems (GIS) to support demographic and economic analyses formed a leitmotif in the papers as speakers described merging and layering diverse sources of demographic and social data on GIS scaffolds. This method yields a better understanding of penetration rates in a spatial context and the geographic impacts of policy measures.

---

[10] Horrigan, John B. and Aaron Smith, Home Broadband Adoption 2007. Pew Internet & American Life Project. Available online at: http://www.pewinternet.org/PPF/r/217/report_display.asp

[11] D. Clayton Brown, Electricity for Rural America: The Fight for the REA (Contributions in Economics and Economic History) No. 29. (Westport, Conn.: Greenwood Press), pp. 3-5.

Grubesic has explored the intersection of geography and infrastructure and described his research on levels of broadband availability across urban, suburban, exurban, rural, and remote communities. Is there evidence of discriminatory practices by private providers, he asked, and does this behavior vary geographically? How can federal, state, and local policies motivate or deter the rollout of advanced services? Data problems constrain the ability to answer these questions. From a methodological perspective, when he looked at the zip code, which is a fundamental unit for the FCC and the Census Bureau, Grubesic found serious inconsistencies in definitions used by FCC and the Census Bureau, a point also made by Flamm and Prieger. The integration of the two datasets is critical, because layering them should allow researchers to take the demographic information from the Census Bureau and overlay it on the infrastructure information on the availability of high speed lines compiled by the FCC. This nexus between datasets turns out to be highly problematic because of inconsistencies in the definitions, and Grubesic concluded, "Spatial statistical approaches are highly susceptible to error when using zip codes as the primary unit of analysis."

Zip codes are not the only spatial measure that researchers employ. Chaudhuri identified problems with definitions of schools districts that inhibited the ability to track public funding. David Gabel of the City University of New York warned that commonly used terms like "urban" and "rural" can be misleading and spatial units may be inconsistently defined or poorly understood. The Metropolitan Statistical Area (MSA), an entity defined by the Office of Management and Budget, is often associated with an urban area and a non-MSA area with a rural area. In its formal definition, "an MSA comprises the central county or counties containing the core, plus adjacent outlying counties having a high degree of social and economic integration with the central county as measured through commuting." Thus, revisions to the MSAs result in new MSAs, removal of MSAs, as well as revisions to existing MSAs."[12] The term actually captures the notion of a social subdivision as much as a spatial unit, and as Gabel pointed out, there are often very large urban areas within an MSA. Since the physical definition changes with shifts in population, comparisons over time are difficult, particularly for environments like Washington, D.C., New Orleans, Atlanta or Los Angeles, where development may occur rapidly or even precipitously.

Culture and users' environment

Complex cultural preferences may also affect users' motivations to adopt technology. Sharon Strover and her students and colleagues at the Telecommunications and Information Policy Institute (TIPI) at the University of Texas at Austin have studied the cultural context of adoption in rural Texas. Strover's research team combines GIS infrastructure maps purchased from private vendors, public data from several statistical agencies, ethnographic interviews and surveys that are mailed to prospective participants, followed by face-to-face interviews as needed. This strategy supports rich but focused analysis and also reveals bias in the data, resulting, in part, from data collection methods. Language and literacy skills, most notably, affect response rates. In their studies for the U.S. Department of Agriculture of sites in Michigan, Kentucky, Texas, roughly half (50%) of the population is Spanish-speaking. The Texas sites in the southern part of the state are nearly entirely Hispanic. Other well known but not insurmountable biases are suspicion of government (or a survey that is believed to be a government document), and race and gender, which affect not only who responds but what is said. The payoff to addressing these

[12] Office of Federal Housing Enterprise Oversight, House Price Index, March 15, 2007, http://www.ofheo.gov/HPIMSA.asp (this site doesn't work)

research challenges are great, Strover noted, as it yields a deep understanding of broadband's impact on a community.

In expanding on the notion of context, Tongia emphasized the users' "setting." He noted the need for researchers to understand the impact of broadband by focusing not just on the infrastructure in the house or neighborhood, but also the devices and software applications they use. There is likely to be a lot of variation in that regard and not all users will have the most current software and hardware at their disposal.

During the first phase of deployment in the 1990s, Tongia continued, many users were introduced to the internet either through institutions of higher education or at work. The next major development was access at home, first through dial-up and then usually through DSL or cable modem to achieve broadband service, although other types of connections are available.[13] Network engineers often talk about "fat pipes" as a colorful shorthand for fast or high-speed transmission. Tongia's point is that a fat pipe to the corner is of little use, if only a pipette links the corner to the home or business. An old machine between the user and the website may compromise the user experience. Still, he commented, many consumers think DSL is slower than cable. Is that the truth, he asked, noting a distinction between a computer architecture and a commercial "marketecture"? And given the perception, how does that view impact competition? So there are two potentially conflated issues: what is out there? And what do ordinary people think is out there?

In engineering terms, answering Tongia's questions requires information about the quality of long haul, second mile, and last-mile services and about end users' usage patterns to appropriately measure people's quality of service experienced. But these engineering measures are actually not well defined even among computer scientists. kc claffy of CAIDA, the Cooperative Association for Internet Data Analysis (http://www.caida.org/home/), which provides tools and analyses promoting the engineering and maintenance of the internet infrastructure including studies of internet traffic, described the "state of internet science" as "abysmal." A well-known advocate for the importance of internet metrics, she listed several reasons for the paucity of data: no federal agency, including the National Science Foundation, is charged with this kind of data collection, and neither the providers, the users, nor the software developers have an incentive to collect or share the data. Yet the potential implications of traffic analysis are significant, as Tongia's comments imply. They inform consumers' mental model of what works and how well it works and therefore frame consumers' decisions about what services to acquire and from whom.

Access to faster connections is one major change for consumers. A second is the expansion of wireless connectivity that enables individuals to connect without a wire tethering them to a specific location and thus moves one step closer to the vision of ubiquitous connectivity anywhere, anytime. Horrigan's recent research for the Pew Internet & American Life Project found that roughly one-third of internet users accessed the internet with a wireless connection from home, work, or "some place else."[14] Public libraries and schools have long been third places after home and work, and the E-Rate Program was explicitly designed to foster expansion of

[13] Some 90% of home broadband connections are either through DSL or cable modems. See U.S. Federal Communications Commission, High-Speed Services for Internet Access: Status as of June 30, 2006 (Washington, DC, February 2007), Table 1, Chart 1. http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-270128A1.pdf.

[14] John Horrigan, Wireless Access, February 2007, Pew Internet and American Life Project, p. 1 http://www.pewinternet.org/pdfs/PIP_Wireless.Use.pdf

access to the internet. In research completed since the workshop, Horrigan has documented ways in which access to wireless appears to signal deeper engagement with cyberspace. Wireless users check e-mail more frequently than other users and are also more likely to get their news online.

Wireless users also typically have broadband at home. Some 80% of wireless users have broadband connections at home, implying that wireless "represents a different quality of online behavior," motivated perhaps by requirements at work or perhaps because it is easy with a personal digital device or a lightweight notebook computer. It is telling that 27 percent of adult internet users access the internet by wireless from some third place other than home or work in contrast to the 20 percent who use wireless at home and 17 percent who use wireless at work.[15]

WiFi "hotspots," like coffee shops and waiting areas of airport terminals, are among these new third places. They are clearly important not only as another means of access but also because of their significance in the emergence of new behaviors and new relationships between people, information, and technology. At the workshop, Martha Fuentes-Bautista and Nobuya Inagaki at the University of Texas at Austin reported on their research into the social use of WiFi in public places. Their work focuses on three major issues:

1. understanding broadband use as a continuous experience
2. how and why people get access from different places and platforms; and
3. how public wireless becomes the entry point for broadband use among the disconnected.

Fuentes-Bautista and Inagaki rely on interviews with employees as well as patrons of such venues to investigate these issues. But they have found it difficult to identify a comprehensive list of establishments from which to develop a sample for survey work. Corporate representatives or providers can be reluctant to disclose information that might be deemed proprietary. Patrons may represent relatively narrow segments of the population, and their willingness to participate may be inhibited by concerns about privacy and confidentiality of the data.

Data confidentiality is not only of concern to individuals who may see a threat to their privacy. Information has commercial value and companies guard their proprietary data as important assets. Richard Clarke, Director of Economic Analysis at AT&T, was one of several speakers who offered the industry perspective. He warned that information on penetration rates is extremely sensitive. In addition, it is important to measure quality of service (QoS) "appropriately," because of the implications that a perceived advantage may have on the market shares of competing service providers. Charles White, Vice President of TNS Telecom, a major telecom market information company, agreed that the data that companies have is expensive to collect; before sharing it, commercial interests need to know how the data will be used.

Data we have and data we need

This section discusses in greater detail some of the major datasets identified by the speakers and their limitations. Problems with these datasets fall into two principal categories: inappropriate and inconsistent definitions; and limitations, bias, and error arising from multiple sources. These problems result in misuse of terms and affect coverage, availability, and reliability of the data and hence potentially undermine subsequent analyses.

[15] Ibid., pp. 1-2

Definitions

All of the speakers noted problems with the definition of broadband. Those who used zip codes in their analysis also identified inconsistencies between the FCC and the Census Bureau in the definition of zip codes. Other speakers, notably Chaudhuri and Gabel, pointed to inconsistencies in definitions in data sets that are used less widely.

Broadband. Under Section 706 of the Telecommunications Act of 1996, the FCC collects standardized information from qualified broadband providers on whether they have one or more lines in service in each zip code.[16] Providers supply the information via Form 477 twice a year at six month intervals and the agency publishes a report and, until recently, made the aggregated data available in downloadable Excel files. More recently, the FCC has decided to make the data available only in the form of Acrobat pdf files which are much more difficult to integrate into statistical analyses packages. The FCC reports the number of carriers with one or more lines in service in each zip code that has at least one provider, but if the number of providers is less than three, the FCC reports an asterisk. Because most communities have at most two facilities-based providers (copper telephone lines supporting DSL service and coaxial television cables supporting cable modem service), this mode of reporting severely limits the usefulness of the FCC data for analyzing the extent of competition available in local markets. The only data that the FCC reports on the number of lines in service are at the aggregate state level, limiting the ability to use the data to study the effect of differing penetration rates by community.

It is true that individual states also collect data, some of it more granular than the FCC's. For example, Gabel pointed to data in Vermont that show where cable networks are deployed throughout the state, arguing that investigators should use state as well as federal data. Unfortunately, coverage is inconsistent from state to state, and in contrast to Vermont, one audience participant said data for the state of Pennsylvania are either "outdated" or represent "projections for where they want to get to." Since the workshop was held, the ConnectKentucky initiative (http://www.connectkentucky.org/) has gained widespread currency as a model for state mapping and data collection. At the workshop, Brian Mefford of ConnectKentucky talked about this public-private partnership to identify gaps and encourage infrastructure build-out in Kentucky.[17]

The FCC did not initially employ the term broadband in its documents. Instead, it defined two levels of service: high speed lines, meaning lines or wireless channels capable of transmitting at rates greater than or equal to 200Kbps in one direction; and advanced services lines, meaning lines or wireless channels capable of transmitting at rates greater than or equal to 200Kbps in both directions. The definition of transmission speed dates to the FCC's first semi-annual report, published in January 1999.[18] At that time, the 200Kbps metric was approximately four times the speed of the typical dial-up connection of 50Kbps and was slightly faster than the 128Kbps rate

[16] The FCC collects data in all 50 states, as well as the District of Columbia and U.S. possessions.

[17] See "Wiring Rural America," *The Economist*, September 13, 2007 for more on ConnectKentucky. Available online at: http://www.economist.com/world/na/displaystory.cfm?story_id=9803963

[18] U.S. Federal Communications Commission Inquiry Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996, CC Docket No. 98-146, January 28, 1999, p. 20, http://www.fcc.gov/broadband/706.html
U.S. Federal Communications Commission, p. 20. (page number listed twice)

of ISDN services, thereby ensuring that ISDN services would not be counted as broadband services.

Prior to 2004, providers with fewer than 250 high speed lines or wireless channels in a given state were exempt from the reporting requirement, thus potentially under-representing rural areas with low population densities. Beginning in 2005, providers below the reporting threshold were obligated to submit Form 477 information. Given the reporting and publication lag, this resulted in a substantial one-time jump in the number of holding companies and unaffiliated entities providing broadband for the period December 31, 2004 to June 30, 2005. Improving the granularity of coverage is a welcome development but it inhibits longitudinal use of the data, since generalizations about low coverage environs are clearly suspect for the period prior to 2004 while conclusions about coverage in areas with better coverage may well be overstated. Moreover, as Sharon Gillett and her colleagues observed, over half of the zip codes in their panel study already had broadband by 1999, so the scope of the data collection precludes investigations of the places which first had broadband available, at least through this source of data.

The chronological scope of availability prior to 1999 is an artifact of the program, and investigators are compelled to seek other sources of information for deployment prior to that time. However, other dimensions of the data collection effort, most importantly the issue of threshold transmission rates, can be adjusted to reflect changing realities. In 2006, the Commission collected more finely grained information about services offered in excess of 200Kbps.[19] Not surprisingly, almost 60 percent of the high speed lines fell into the category of greater than or equal to 2.5Mbps and less than or equal to 10Mbps, and just under 5 percent had transfer rates greater than or equal to 10Mbps.[20] As a number of speakers noted, efforts to refine the definition of broadband to reflect the changing nature of broadband services and the availability of ever-higher-data rates is long overdue. And indeed, FCC Chairman Martin announced in his testimony before Committee on Energy and Commerce in the U.S. House of Representatives, that in the Fifth Inquiry, the commission seeks comment on whether the term "advanced services" should be redefined to require higher minimum speeds.[21]

Zip codes. In his testimony, Chairman Martin also cited a proposal put forward in September 2006 to improve data collection be examining specific geographic areas and integration of FCC data with data collected by states and other public sources. Workshop participants acknowledged the importance of integrating data from state and federal sources, but more forcefully drove home the problems with using zip codes.

Both the FCC and the Census Bureau use the zip code as a unit of analysis but they define it differently, creating problems when researchers seek to merge data sets. The Census Bureau has created new statistical entities called "zip code tabulation areas" (ZCTAs) which represent the generalized boundaries of US Postal Service zip code service areas; these are not equivalent to the older zip codes, and it is clear that the FCC is **not** using the ZCTAs. Moreover, not all zip

[19] Written Statement of the Honorable Kevin J. Martin, Chairman of the Federal Communications Commission before the Committee on Energy and Commerce, U.S. House of Representatives, March 14, 2007 http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-271486A1.pdf%20--%20see%20pages%203-4, p. 4

[20] U.S. Federal Communications Commission, High-Speed Services for Internet Access: Status as of June 30, 2006 (Washington, DC, February 2007), Table 1, Chart 1. http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-270128A1.pdf , Chart 9.

[21] Martin, Testimony, 2007, p. 4

codes denote spatial units although they are widely believed to do so. Rather, zip codes reflect their origins in the postal service and are actually linear features, corresponding to mailing addresses and streets in USPS service areas. Finally, zip codes are thought to denote relatively small spatial units, at least in comparison with states and counties.

Zip codes do represent relatively compact areas in urban environments, Grubesic found, but not in exurban or rural areas. Flamm agreed that zip codes are actually fairly coarse and can measure relatively large territories. In addition, Flamm noted that the FCC uses a publicly undocumented and proprietary database, which the agency purchases from a provider. The provider, however, uses zip code mapping software that adds and drops zip codes relatively rapidly, further complicating the ability to figure out what territory corresponds to what code over time.[22]

A second set of problems arises from the way that the FCC collects, reports, and aggregates – or does not aggregate – data within zip codes. First, prior to data collected in June 2005, no zeros are measured, Flamm found, so identifying zip codes' areas without any service requires subtracting all of the zip codes where broadband service is provided from the universe of zip codes used by the FCC for this period. Unfortunately, the FCC has chosen not to publish a list of the universe of zip codes used in its classifications. Like the expansion to include smaller providers, though, this welcome change in data collection inhibits longitudinal studies.

Second, zip codes where 1 to 3 providers exist are reported at the categorical level. Actual counts are provided for zip codes with more than 3 providers, in effect masking under provisioning while detailing better resourced areas where the local market is presumably more competitive. Prieger echoed this point, observing that researchers cannot say anything about monopoly vs. duopoly vs. oligopoly in areas of served by 1-3 providers.

Third, Flamm argued, the cartographic presentation can be misleading, especially in areas where zip codes are large, which are generally the rural areas, since the value, whether categorical or an actual count, is then mapped over the entire expanse.

Thus, three factors converge to potentially overestimate coverage, or to give the appearance of overestimating coverage, in what may actually be under-resourced areas: the absence of the zero measurement prior to 2005, the categorical representation of the service areas in which 1-3 providers are present, and the coarse geographic measure associated with zip codes in rural areas. Flamm offers two examples:

- In the 78731 zip code in Austin, Texas, where he resides – a city that is generally considered a center of innovation – the FCC statistics indicate 24 broadband service providers. "After an exhaustive search for alternatives, however, I know that there is only one physical, local broadband service provider available within my immediate neighborhood."[23] Residents of affluent Fairfax County in Northern Virginia had precisely the same experience of limited availability of connectivity to the residential neighborhoods at the same time that it contained one of the backbone hubs.

- More generally, he cited research done by the GAO to assess the number of competitive entities providing service nationally. Their study showed that the median number of providers serving households at the zip code level went from 8 to 2. In Kentucky, for

[22] Flamm 2007, p. 5

[23] Flamm, 2007, p. 8

example, where the GAO's calculation based on FCC information showed that service was available to 96 percent of the state's population, ConnectKentucky did a more sophisticated analysis and found that only 77 percent of the state's households had access to broadband service.

Flamm also took issue with the way providers are defined and identified. The FCC defines providers as "facilities based providers" of high-speed services to end user locations anywhere in a state in which they own hardware facilities. These, he points out, can be service providers, not actual infrastructure providers or "hardware pipe provisioners," in some local markets, but not in others within a given state. It is unclear whether or not such mixed providers of broadband service distinguish carefully between zip codes in which they own some of the hardware used to provide the service, and zip codes in which their service is branded and resold by the actual hardware pipe provider. If multiple providers "brand the same pipe," the view of competition is affected. Further, since identities of providers are not known, Prieger added, nothing can be said about industry dynamics (entry and exit), impact of multi-market contact, or intermodal competition (for example, cable vs. DSL). In addition, the FCC identifies location of service by the location to which the bill is delivered. If the bill goes one place (say, a post office box), and the line is actually installed elsewhere (say at a home), then the companies are reporting the zip code of the post office box, not the home, to the FCC. Thus, the FCC data measures what zip codes broadband recipients are being billed in, not necessarily where the service is available or actually consumed.

Available data and their limitations, bias and error

The data assembled by the FCC is intended to assess competition by looking at what providers have deployed. It is not geared to evaluate quality of service, performance at the desktop, or even penetration. Even though it does provide information on availability, it does so imperfectly. Penetration rates, which are necessary for analyses of regional economic impacts, are usually developed by layering demographic data, typically from the Census Bureau, over geographic data. That produces problems with definitions of zip codes which inhibits merging the two data sets. Flamm has devised a strategy for reconciling the discrepancies between the two zip code definitions but acknowledged that the resulting sample would under-represent remote, sparsely populated rural areas. The strategy itself involves limiting the analysis to the intersection of zip codes that appear in both the FCC and Census Bureau pools. Two types of zip codes, "geo" zip codes and "point" zip codes, are assigned and the boundaries laid out, allowing for the fact that the spatial units associated with the two schemes will probably not coincide perfectly but will cover a similar area. On this basis, he has been able to construct a database that allows him to examine spatial, demographic and topographic as well as economic variables.[24]

Other than the FCC, workshop participants described several other sources of data, namely the Bureau of Labor Statistics, Bureau of Economic Analysis, and Census Bureau. The Census Bureau conducts the national decennial survey as well as more focused studies at more frequent intervals, notably the Current Population Survey, the American Communities Survey, and the Current Employment Statistics Survey. The advantages of using these collections are long runs of data, broad coverage at a national scale, quantity and variety, and the professionalism and prestige of the federal statistical agencies. The Bureau of Economic Analysis (BEA) and the Bureau of Labor Statistics (BLS) are two major sources of economic and industry-related data. The E-Stats program is a comparatively recent and also employs separate Census Bureau surveys

[24] Flamm's strategy and initial conclusions are described in Flamm, 2007, pp. 9-10.

to compile information on economic and e-commerce activities. In general, these agencies are slow to adopt methodological changes, but they adjust the categories of information they collect on special topics.

Triplett co-authored a paper with his colleague Barry Bosworth at the Brookings Institution in 2003 in which they detailed some of the recent progress in data collection at these agencies as well as some then still-remaining issues, notably inconsistent data sources that affect measures of productivity.[25] Greenstein notes that BEA has recently begun publishing information on investment levels in information/communications technologies by industry and has included wages and salaries for workers in some locales. However, these studies address information technologies at a broad level; internet and other components are not isolated in the research. Somewhat bravely, the Census Bureau attempted a survey of software used by firms for the year 2003, Greenstein comments, but the task proved monumental and the results were inconclusive. The design called for surveying firms, not individual establishments, and the practical issues were daunting, starting with deciding who to contact and what information, presumably in the form of an inventory of software tools, even existed.

In general, researchers have trouble finding data at a suitably granular level. This is a problem, for example, that affects studies of firms, salaries and wages, and pricing. One solution is using private sources of information, but these are expensive and can be limited by issues of confidentiality and proprietorship. Greenstein, Forman, and others have made extensive use of data supplied by the business intelligence and marketing company Harte-Hanks. Not only are the datasets are expensive, but they naturally reflect the interests of the company's clients who have paid for the initial surveys. Thus, the content of the data files is geared toward marketing and not necessarily toward the questions that investigators may have. In a subsequent e-mail exchange, Greenstein has offered several illustrations:

- The company currently archives its data but has no plans for donating it to a public (or even private) curatorial facility, potentially inhibiting developing longitudinal studies of change over time. Moreover, digital data is fragile and requires active management. A corporate policy of benign neglect, which would not necessarily destroy physical records, can result in unusable digital data, which effectively destroys it.

- The company focuses on coverage of potential sales targets. This strategy overlaps with the statisticians' need to get the biggest users. But it also means it will be deficient in some systematic ways. Greenstein and his colleagues have done comparisons against county business patterns and have noticed the following: the coverage of small firms is rather inadequate. The reasons are obvious. Unlike a government agency concerned about the need to be complete, there is no systematic over-sampling of the underrepresented population. (Sampling issues matter to statisticians but not to most clients.)

- Harte-Hanks data provide a measure of the number of programmers, but they do not provide detail on the composition of the computer workforce – their quality, education, experience, wages, or degree of autonomy. Without such measures, it is not possible to get at the core issues about whether the degree of decentralization is changing over time, whether computing is affiliated with biased technical change, and so on.

[25] Bosworth and Triplett, pp. 31-35

For demographic research, the key federal data source is the Current Population Survey, which is a monthly household survey conducted by the Bureau of the Census for the Bureau of Labor Statistics, which has included questions about computer and internet use in households in the period 1994-2003. This data collection effort was the basis for the series of reports by the National Telecommunications and Information Administration (NTIA) on computer use beginning in 1994; internet use was added in 1997. The program was ended after 2003 but the data can still be downloaded from the Bureau of Labor Statistics' website: http://www.bls.census.gov/cps/computer/computer.htm.

Federal surveys of computer and internet access and use provide baseline information according to basic demographic characteristics (age, location, educational achievement, and income). Little is known from these federal surveys about behavior, choices and motivation, or even what people pay for broadband service, although many of these topics are explored in surveys conducted by the Pew Internet Project. Social scientists typically augment these large, primarily descriptive studies with information from private sources that they either purchase or create through new data collection efforts. In some cases, though, the data may contain outright errors as Strover discovered in her use of GIS data obtained from private sources. It also may be, as Flamm said that the methodology for collecting and categorizing data is not documented.

Compared with the national statistical efforts, academic studies are focused but small so that the breadth of the national surveys is balanced by the depth of the academic surveys. The kinds of databases that Flamm, Grubesic and Goldfarb describe are time consuming and expensive to build and tend to be geographically restricted so that the detailed work is needed to resolve the discrepancies and establish correct linkages. Computer scientists call this "registering" the data and it means the techniques required to achieve valid integration of disparate data sets. It is a problem that is endemic to use of multiple quantitative datasets that, when assembled, can yield wonderful information but in themselves are more heterogeneous than they appear. Decisions, like Flamm's resolution of zip codes, are always part of the research process, so that documenting the management of the data becomes an integral component of presenting results since it may well happen that the process of integrating the datasets introduces a bias, as Flamm and others readily acknowledge.

The work done by Strover and her students Fuentes-Bautista and Inagaki reflects a range of survey designs, sampling techniques, and methods of information capture, including telephone interviews, face-to-face interviews and mailed surveys. Sometimes formal sampling is not possible, as shown in the research done by Fuentes-Bautista and Inagaki. Their project focuses on an intrinsically self-organizing population, patrons of commercial establishments that have WiFI, making the sample opportunistic, what statisticians call a "non-probability sample." These mainly qualitative research methods produced nuanced portraits of special populations, but, by their very nature, do not permit generalizations about the general population.

These projects all have methodological limitations: low response rates, strained interpersonal dynamics, concerns about personal privacy and inappropriate disclosure of information, suspicion, and reliability. Fuentes-Bautista and Inagaki attempted to correct some of the bias in their design by creating a database of available hotspots, but they could not find a "comprehensible" list of them. Strover explained ways in which she and her colleagues remedy low response rates, particularly from groups who tend to be under-reported. Follow-up calls, mailed as well as door-to-door surveys, and use of bilingual interviewers are among the methods. Still, self-reporting, whether by phone or in person, can be biased by a number of factors, as Strover and her colleagues documented in their cross cultural research, including language, gender, age, and perception of status. Roles also introduce constraints between otherwise similar

interviewer and interviewee; corporate representatives may be reluctant to provide the information because it is confidential or proprietary, or, as the Census Bureau found in its ambitious but ill-fated attempt to survey companies for software use, the individuals answering the question simply may not know the answer.[26]

Goldfarb points out that clickstream data, which is collected at the device, may provide a useful way to offset questions about reliability of self-reported information. This type of research has a relatively long history in the computer science/information science community, which has collectively devised multiple experiments that combine analysis of computer log files with real-time observation and interviews. Log files or the more granular clickstream data show what people actually do online. "It's not always pretty," Goldfarb said, but "it provides rich detail." He also acknowledged that clickstream data is hard to use. Using it entails more manipulation, known as "data cleaning," and analyzing the "cleaned" data requires more statistical sophistication.

As computer and information scientists have learned, though, collecting this kind of information sparks concerns about personal privacy. Formal experiments in laboratory settings are bounded by strict policies governing data management and use. The kinds of concerns about sample size and composition can also arise in small, laboratory-based projects. Thus, corporate data collections are appealing because of their scale and scope. According to Goldfarb, two private companies, comScore and Nielsen//NetRatings, collect this data on home usage but neither shares the raw data with researchers. Concerns about privacy and disclosure are not attached solely to personal information. As Fuentes-Bautista and Inagaki discovered and Wallsten reiterated, companies are careful about the information they provide, particularly if the information reflects or might reflect upon human, corporate, or computational performance. This enhances the significance of the data on internet traffic that CAIDA collects.

Conclusions and recommendations

Any research undertaking in the social sciences must confront the following questions: What is the evidence on which you base your conclusions? How reliable is that evidence? These questions permeated the papers and discussions at the workshop. As has been noted, speakers and participants concluded that the existing datasets all have limitations that inhibit their use. Some of these limitations render the data almost meaningless for some questions; others require creative work-arounds. The major issues include the following:

- The large, national level data sets do not provide data at a suitably granular level of detail. They support broad generalizations at the national level but they are not well-suited for:
    - Studying closely the behavior of firms;
    - Refining models of workforce participation, incentives, and rewards;
    - Understanding user adoption at local or regional levels.

- Existing data sources data provide limited insight into the extent of consumer choice of broadband service providers. They also overstate the availability of broadband services because of the way in which they are reported.

[26] Greenstein, 2007, p. 9

- The definition of broadband articulated by the FCC is out of date and should be revised to capture current conditions. Accompanying that effort, the FCC should collect data in such a way so that researchers can develop an adequate picture of how the speed of broadband services is evolving.

- Zip codes are inconsistently defined by the two major agencies, the FCC and the Census Bureau, inhibiting merging datasets from multiple sources, necessary to obtain richer analysis. Other units of analysis, for example, the MSA and the school district, suffer from similar inconsistencies and misunderstandings. Solutions have been proposed to work around the inconsistencies, but nonetheless there remain severe limitations on what questions the data can support.

- Data measurements and even appropriate metrics to measure the quality of service experienced by end-users are lacking.

- Data on broadband pricing and business and household expenditures on broadband services are not readily available except in aggregate. Even here, meaningful categories of price data are not available in useful form.

- Data on broadband traffic patterns and its evolution over time are also not readily available. The appropriate metrics for analyzing such traffic and addressing such questions as the patterns of internet interconnections and routing policies are not well-defined, even within the technical community.

- Data that may be available from the private sector for business or residential users can be expensive, opaque, and even erroneous. The data provider may place restrictions of use of the data, as evidenced by the two major providers of clickstream data, comScore and Nielsen//NetRatings. Moreover, there is no guarantee that the owner will archive the digital data properly so that they are available for long term use.

- Privately funded academic research does compensate for some of the deficiencies in the large datasets, but these studies tend to be customized, focused and hence not necessarily comparable. Here, too, the data are not necessarily archived. Researchers have displayed varying levels of reluctance to deposit and reuse data which also can inhibit comparative studies as well as verify the analysis."[27]

A recurring theme among researchers participating in the workshop is the importance of the local. Broadband connections are generally localized in a specific geographic area (i.e., wired connections are tethered to a specific location and wireless services are generally short-range). Significant differences in the availability and quality of service may occur over relatively short geographic distances, issues which are important to a full understanding of broadband markets. Moreover, the notion of quality of service might well be broadened beyond traditional metric of bandwidth to include assessment of the users' experience to allow addressing such questions as "Are there measurable effects of improvements in user interface design?"[28] or "Does security

[27] Reluctance to share data by scientists and social scientists is documented in a report by the Association of Research Libraries, "To Stand the Test of Time: Long Term Curation and Management of Large Data Sets in Science and Engineering; A Report from the Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe." September 26-27, 2006, Arlington, Virginia.

[28] Greenstein has raised this question in the context of the use of the Consumer Price Index, which he notes is a transactional index, which accurately reflects the changes to prices at which users transact for Internet

mean anything to end users and with the rise of identify theft, are they willing to pay for it?" With workshop participants repeatedly returning to the user and the context of use, a critical need is to collect data on broadband that are geographically disaggregated in order to understand the local impacts of broadband.

As the internet matures beyond a digital communications system to an infrastructure that supports a range of applications and services, we can expect the problems of measurement to become even more complex and difficult. At the same time that more granular data is needed, the landscape is becoming more intricate. More competitors are entering the market, offering different combinations of services for both infrastructure providers and end-users.

Finally, the workshop called for multidisciplinary and interdisciplinary approaches, engaging engineers who understand implications of changing technologies for users, economists with a broad variety of expertise, sociologists versed a range of skills from language to statistics, and industry participants with access to important datasets and a framework that enables them to share the information to advance the state of knowledge without incurring liabilities.

Acknowledgements

This essay is based upon a day-long workshop organized by Kenneth Flamm of the University of Texas at Austin, John B. Horrigan of the Pew Internet & American Life Project, William Lehr of the Massachusetts Institute of Technology, and Sharon Gillett of MIT (at the time of the workshop).[29] The text was written by Amy Friedlander, in collaboration with Kenneth Flamm, John Horrigan, and William Lehr. Elizabeth Maida, a student at MIT, did an outstanding job of transcribing tapes of the day's proceedings. In the course of writing the essay, Dr. Friedlander benefited from conversations and email exchanges with Shane Greenstein and Myron Gutmann.

This essay should be cited as Kenneth Flamm, Friedlander, Amy, Horrigan, John B., Lehr, William. *Measuring Broadband: Improving Communications Policymaking through Better Data Collection*. (Washington, D.C.: Pew Internet & American Life Project, 2007). Available online at: www.pewinternet.org

---

access" and not a utility-based index, which would attempt to capture the value to users of a new technology. See Shane Greenstein, Is the Price Right? The CPI for Internet Access. A Report for the Bureau of Economic Analysis, December 20, 2002.

[29] Gillett is now head of the Department of Telecommunications and Cable for the Commonwealth of Massachusetts.

Appendix

Measuring Broadband: Problems and Possibilities
A Workshop at the DC Office of
The Pew Research Center
1615 L St St. NW, 7th Floor
Washington DC
June 28, 2006

Co-sponsored by
Pew Internet & American Life Project
University of Texas at Austin, with support from the National Science Foundation
Massachusetts Institute of Technology

Communications infrastructure plays an increasingly important role in our society. It is a critical infrastructure relied upon to run public utilities, transportation, and security systems. In a global economy, all manner of enterprises rely on communications infrastructure to deliver new and innovative services to customers, coordinate production, and reduce transaction costs. On a social level, the internet's interactive nature allows people to create, consume, and exchange a wide range of information, which fosters social connectedness and helps build social capital. At the household level, expenditures on services delivered over broadband may soon exceed what formerly were expenditures on telephone and cable television services, and account for a major slice of consumer spending. Measurement and monitoring of changes like these will be critical to our nation's understanding of its economic and societal health. Yet accurate understanding depends on data that is not now collected in consistent and predictable ways.

In the United States, there is an explicit national policy goal to have competitive and affordable high-speed internet service widely available to Americans by 2007. But techniques currently used to measure broadband infrastructure and user adoption, which would help monitor progress toward that goal, have limitations. Current federal reporting requirements do not sufficiently measure deployment of high-speed infrastructure; state and local requirements are piecemeal and inconsistent. Virtually no detailed and scientifically collected data are available on critical economic variables, like pricing and measures of quality of service. Although private surveys presently do an adequate job of measuring the broad contours of users' broadband adoption and online behavior, exclusive reliance on this approach may become problematic as networks evolve, services are increasingly differentiated, and connection speeds vary by vendor or class of service.

This one-day, invitation-only workshop will gather leading academics, government statisticians, and practitioners in the private and non-profit sectors who have conducted research or gathered empirical data in this field.[30] The objectives of the workshop are to identify gaps and shortcomings in current measurement techniques and propose improvements. Participants will be drawn from universities and public policy research organizations that have worked on measurement of broadband deployment and use, and from government organizations and statistical agencies collecting data related to broadband.

The goals of the workshop will be to:

[30] This workshop is a follow-on activity to a June 2005 University of Texas at Austin-organized workshop on "Internet Use in the Americas," supported by the NSF.

a) Develop recommendations on how to improve data collection on the communications infrastructure, both its scope and economic and social impact.
b) Stimulate research and experimentation among researchers in the public and private sectors that will result in policy relevant research on economic and social dimensions of the "broadband society."

To the extent that the workshop's recommendations are aimed at government agencies, the agencies most interested in the workshop are likely to be the Bureau of Economic Analysis (BEA) in the U.S. Department of Commerce, the Bureau of Labor Statistics (BLS) in the Labor Department, and the Federal Communications Commission.

AGENDA

8:00 – 8:30 AM
Continental Breakfast

8:30 – 8:50 AM
Introduction: Workshop Organizers
Kenneth Flamm, University of Texas
Sharon Gillett, Massachusetts Institute of Technology
John Horrigan, Pew Internet & American Life Project

Overview and Framing: Jack Triplett, Brookings Institution

8:50 – 10:15 AM
Panel: Perspectives on Economic Research Using Broadband-Related Data
Moderator: Charles R. Hulten, University of Maryland
Panelists will briefly outline the purpose of their research, the data they used, and the challenges they encountered. Panel discussion will focus on changes and additions to current data collections that would significantly improve economic research related to broadband deployment and use.
Panelists:
Anindya Chaudhuri, University of Texas
Kenneth Flamm, University of Texas
Chris Forman, Carnegie Mellon University and Shane Greenstein, Northwestern University
Avi Goldfarb, University of Toronto
John Horrigan, Pew Internet & American Life Project
William Lehr and Sharon Gillett, Massachusetts Institute of Technology
James Prieger, University of California Davis
Gregory Rosston, Stanford University
Scott Wallsten, American Enterprise Institute

10:15 – 10:30 AM
Coffee Break

10:30 – 11:30 AM
Panel : Other Research Perspectives
Moderator: Shane Greenstein, Northwestern University
Panelists from engineering and non-economic social science disciplines will complement the previous panel's perspectives on research purpose, data, challenges, and improvements.
Panelists:
KC Claffy and Tom Vest, University of California San Diego (CAIDA)
Martha Fuentes-Bautista and Nobuya Inagaki, University of Texas

Amy Glasmeier, Pennsylvania State University
Tony Grubesic, University of Cincinnati
Judith Mariscal, CIDE, Ciencias Sociales
Jorge Schement, Pennsylvania State University
Sharon Strover, University of Texas
Rahul Tongia, Carnegie Mellon University

11:30 AM – 12:15 PM
Panel: Private Sector Perspectives
Moderator: David Young, Verizon
Industry participants will provide brief overviews of their research needs, the data they use and the problems they encounter. Panel discussion will focus on firms as both users and producers of broadband-related data, and dealing with competitive sensitivities.
Panelists:
Richard Clarke, AT&T
Roman Krzanowski, Verizon
Michael Nelson, IBM
Robert Pepper, Cisco
Chuck White, TNS Telecoms
Bill McCready, Knowledge Networks
Discussants:
Derek Turner, Free Press
Frederick Weingarten, American Library Association

12:15 – 1:15 PM
Lunch

1:15 – 2:45 PM
Roundtable: Government Data Collection
Moderator: Jack Triplett, Brookings Institution
Attendees from federal and state government agencies will answer questions about the data they collect, its intended purpose and associated challenges.

2:45 – 3:00 PM
Coffee Break

3:00 – 3:45 PM
Roundtable: Government Policy Perspectives
Moderator: Robert Pepper, Cisco
Selected government attendees will discuss the needs, uses, and challenges for data in broadband-related policy making. Panelists will include Lisa Sutherland, Senate Commerce Committee Staff , Senator Ted Stevens (R-AK), and Colin Crowell, Telecommunications Staff, Rep. Ed Markey (D-MA).

3:45 – 4:45 PM
Roundtable Discussion: An Agenda for Improving Data Collection and Use
Moderators: Workshop Organizers (Flamm, Gillett, and Horrigan)

4:45 – 5:00 PM
Summary and Consensus Recommendations (Flamm, Gillett, and Horrigan)

List of Attendees

Anna	Aizcorbe	U.S. Department. of Commerce
Dennis	Alvord	U.S. Department. of Commerce
BK	Atrostic	U.S. Census Bureau
Kim	Bayard	Federal Reserve Board
Ellen	Burton	Federal Communications Commission
David	Byrne	Federal Reserve Board
Kenneth	Carter	Federal Communications Commission
Anindya	Chaudhuri	University of Texas at Austin
Barbara	Cherry	Federal Communications Commission
KC	Claffy	Cooperative Association for Internet Data Analysis
Richard	Clarke	AT&T
Michael	Clements	Government Accountability Office
Mark	Cooper	Consumer Federation of America
Carol A.	Corrado	The Federal Reserve Board
Colin	Crowell	Office of Rep. Edward J. Markey
William	Ennen	Mass Tech
Kenneth	Flamm	University of Texas at Austin
Christopher	Forman	Carnegie Mellon University
Martha	Fuentes-Bautista	University of Texas at Austin
David	Gabel	Queens College City University of New York
Sharon	Gillett	Massachusetts Institute of Technology
Amy	Glasmeier	Pennsylvania State University
Avi	Goldfarb	University of Toronto
Shane	Greenstein	Northwestern University
Anthony	Grubesic	University of Cincinnati
Michael	Holdway	U.S. Department of Labor
John	Horrigan	Pew Internet & American Life Project
Charles R.	Hulten	University of Maryland
C. Suzanne	Iacono	National Science Foundation
Nobuo	Inagaki	University of Texas at Austin
Sherille	Ismail	Federal Communications Commission
David	Johnson	U.S. Census Bureau
Roman	Krzanowski	Verizon
William	Lehr	Massachusetts Institute of Technology
Judith	Mariscal	CIDE, Mexico
Robert	McCelland	U.S. Bureau of Labor Statistics
James	McConnaughey	U.S. Department. of Commerce
William	McCready	Knowledge Networks
Brian	Mefford	Connect Kentucky
Thomas	Mesenbourg	Bureau of Census
Michael	Nelson	IBM
Stephen D.	Oliner	The Federal Reserve Board
Robert	Pepper	Cisco
Kenneth	Peres	Communications Workers of America
Melissa	Pollak	National Science Foundation
James	Prieger	University of California, Davis
Gregory	Rosston	Stanford University
Jorge	Schement	Pennsylvania State University

Christina	Speck	U.S. Department. of Commerce
Sharon	Strover	University of Texas at Austin
Rahul	Tongia	Carnegie Mellon University
Jack	Triplett	The Brookings Institution
Eric	Van Wambeke	California Public Utilities Commission
Tom	Vest	Cooperative Association for Internet Data Analysis
Scott	Wallsten	Progress and Freedom Foundation
Philip	Webre	U.S. Congressional Budget Office
Chuck	White	TNS Telecoms
Phyllis	White	California Public Utilities Commission
Irene	Wu	Federal Communications Commission
Derek	Turner	Free Press
David	Young	Verizon