Data Constraints & the Internet Economy: Impressions & imprecision.

By Shane Greenstein

NSF/OECD meeting on "Factors Shaping the Future of the Internet."

Comments welcome.

February 21, 2007

Introduction

Today I would like to assess the statistical foundations for addressing the most central economics question in Internet studies: what is the contribution of the Internet to economic growth in the United States over the last decade? What will be its contribution over the next decade?

I will focus on this question for convenience. My underlying motivation is broader. I believe statistics can inform policy debates. I make that statement as an empiricist. I have observed that economic statistics are a foundation for policy formulation in a multitude of areas.

Here is the problem: There are not shortages of Internet policy debates, but there is a shortage of facts. To be more precise: there is a shortage of meaningful data.

When a meaningful statistic is available it quickly becomes central to a debate. I have observed this shortage in many debates, such as those about inequality of access and use of the Internet, about the effect of the Internet on international trade, about the changing requirements for human capital accumulation in information technology markets, and about the spread of electronic government. I could go on. It is a pervasive phenomenon.

If you do not believe me, let me begin with a narrow motivation. The privatization of the Internet – an event that NSF managed more than a decade ago – is one of the greatest gifts to the US and world economy to ever come from the US research community. Wouldn't someone love to say – with some degree of confidence – what the contribution

of the Internet was to the US economy? To be sure, it is not straight forward question to address. One must account for all the incremental changes brought about from investment. Actually, it is even harder than usual (for a new technology), because one must account for all the increases in productivity as well as the wastefulness that occurred during the boom and bust. Still, doesn't that seem like a comparatively important question to address? It turns out: we do not know the answer. More to the point, we do know pieces of it, but many pieces remain unexamined. Worse yet, there is no lively debate about the remaining open questions because scholars do not do not have the data to find out.

This essay has a simple and broad point. The state of statistical information about the information economy in the United States is rather mediocre. There are oases of excellence, but these are rare and not the norm.

What is my benchmark for "mediocre"? I say that primarily in comparison to other industries, such as, for example, cement and concrete. We have marvelous statistical data about thousands of cements and concrete plants throughout the US, as well as the users in many locales. We know a lot about the price of cement and concrete, productivity improvement in cement and concrete, the contribution of these firms to the tax base of their local economy, and, even, how much they contribute to pollution in a locality. It goes on and on. I would conjecture that (if a policy maker cared to know) we can predict how many plants will enter a local region when the US Congress passes a new highway construction bill.

In Internet studies, in contrast, we have little comparable data about the prices, quality, taxes, employment or revenues. The frontier of research is still at a descriptive level because we do not have the data. For example, we would like to know how economic growth changes when there is more broadband competition, but we cannot answer that because we do not know how firms respond to additional competition, or how demanders respond to lower prices and higher quality and more rapid investment. In fact, we are not sure how many cities in the US have competitive broadband markets. (The answer is: we have one statistic that nobody trusts, so nobody knows for sure. More coming below.) According to industry engineers the problems run deep: we also do not know the patterns for the basic flows of traffic across the entire network. Even if we had

good information about the "branches" in the network, we are not sure whether (or how much) it would alter the quality of experience of users in a particular location on the network.

We simply know a lot more about the economics of cement and concrete than about the economics of the Internet. That is because we have data about one market and do not have the data to learn more about the other.

In this essay I am going to dwell on describing the state of data. I hope to describe the range of places where nobody can connect the dots. I hope it will motivate others to initiate projects that will fill gaps piece by piece.

Today I will concentrate on the US situation, and will discuss numerous examples. My impression is that the US situation looks worse than the situation found in other developed countries. This impression comes from reading OECD publications, where some impressive statistics appear from time to time. However, I will not back up that impression with evidence, so I am happy to be corrected, if wrong.

In making these observations I am not being critical of US administrators and statisticians who spend all their working time collecting economic statistics in the US. I know many of these people. In the past, when I could, I have worked with many of them to improve things in incremental ways. They deserve praise for the impossible situation most of them face on a daily basis. I know that many of them have tried to persuade their political bosses to pay for anything other than cement and concrete statistics, but those arguments ran into a brick wall.

Rather, my view is that we are watching an industry much like the railroad industry in 1885. By that point it was obvious to any waking observer that something dramatic was afoot, but putting together the entire picture was quite challenging. Why was it challenging? The problem was basic: the object under study changed quickly, and the problems kept getting worse because nobody documented a base year. As a result, there was not much data for documenting change over time. The community of interested users and data-collectors was largely voluntary in that era, and together they worked to improve, modify, or initiate statistical studies.

More to the point, political actors and research community did not invest time or money in improving data collection then. In that era, as in our own, it must have been

difficult to make informed policy about the changing nature of tax collection in localities and at a national level, about the likelihood of productivity improvements across industries and locations, about changes occurring in inter-state trade, about changes in the location of economic activity, about changes in work force requirements – in other words, about the basic parameters of economic policy.

The review proceeds as follows: I will first provide an overview of the statistical issues in the US, where I know the situation in some detail. Then I will provide a little case study for illustrating issues. This involves explaining some of the issues in the official US price index for Internet access and some of the basic open issues in broadband policy. Throughout I will try to identify places where basic research has high value if it simply documents what is happening in a way that permits comparison over time.

Throughout I will be quoting extensively from three literature reviews, one done by Chris Forman and Avi Goldfarb (2006), one by me, Greenstein (2005), and another by a Jeff Prince and me, Greenstein and Prince (2006). Some readers will also notice some thematic overlap with Griliches' AEA (1994) presidential address about the data constraint in economic measurement.

A warning: When I give this essay the flavor of a light literature review, I do so to support the impression that my review is thorough. At the same time, I intend for this to be a personal essay, not a long literature review that drones on and on. If someone wants the lit review, I expect them to see go find the references in the lit reviews.

Said bluntly, unlike a normal literature review it is not my goal to politely acknowledge each contribution. Rather, my goal is to impolitely identify the gaps. Sometimes I will cite papers and researchers by name, but often I will not be thorough. I make no promise or pretense to cite everyone (and I apologize for leaving some people out). If the reader wishes for more detail, those literature reviews offer a start.

Gaps in the state of knowledge

What type of data would be required to address a basic question about the contribution of ICTs to US economy? Here is what an economic researcher would need to address this fundamental question:

- An ideal data set would measure variation in investment in computing and communications hardware and software, as well as provide a reasonable measurement of labor force employment. That would allow a researcher to identify how much IT matters for firm performance.
    - The Bureau of Economic Analysis now publishes an estimate of investment levels in ICT by industry for some recent years and recently began publishing estimates about the wages and numbers of IT workers in some locales. Frankly, they deserve credit for this. The situation is better than it used to be, and many scholars have taken advantage of it (more below). It also is not ideal, so we should not get carried away.
    - To be sure, this is a large topic and several authors have tried to chip away at the issues. Erik Brynjolfsson and co-authors (Brynjolfsson and Hitt, 2002), as well as Chris Forman, Avi Goldfarb and I, have made an estimate for the use of ICTs or the Internet by business in 2000 (e.g., see Forman, Goldbarb, and Greenstein, 2002). Arora and Forman (2006) have also made estimates of the local software service economies. Mark Doms has several studies that look at estimating PC use in business establishments. The research group at Irvine (Ken Kramer, et al), known as CRITO, also has done some original survey work. This is progress, but it has not resolved the basic open questions (more below).
- It would be ideal to have reasonable measurement of variation in use of broadband services, between industries on average, and between firms within the same industry. This would also require some "controls" for whether the equipment was owned by an establishment or rented from providers for monthly fees, both of which are common. As far as I know, no consistent historical data exist on this. Again, I will say more below.
    - Why do historical data matter? Here "historical" usually means "consistent time series over the last ten years." Something a bit

longer would be nice too, but such a dream is usually impossible in practice. Here is a illustration. When Forman, Goldfarb and I estimated differences in use of the Internet across Industries, we went looking for prior census on IT use across industries. Even something from the mid 1980s would have been fine. We could not find any such list, and only much later did we learn of a project that came close – that is, Ed Wolff's Herculean research project tracing changes in demand for information workers across post WWII industries, using a Machlup framework. Why does this matter? Using our data we could not say, with any degree of confidence in any ballpark, whether the diffusion of the Internet had fundamentally altered the buying patterns for ICTs across the US economy or not. Hence, a fundamental question went unaddressed.

Many economic policy issues would improve with data about the geographic variation across the US economy. Many facets of electronic commerce, such as broadband or contract programming or local-area-network maintenance for hire, are inherently a local good/service. Their direct impact should be local. Some cities have robust markets for such services and some do not. It would be useful to know if/why this variation matters.

- For measuring direct effect on economic growth in a metro area, a researcher would need to know about the availability and (especially) use by business establishments of these technologies.
    - The first research to look at this topic in Internet studies looked at registrations of domain names, including work by Mathew Zook, and by Jed Kolko. Once again, the more recent work by Forman, Goldfarb and myself also tried to catalogue how areas differed. We eventually found there was plenty of good company with work by Ken Flamm, Tony Grubesic, and James Preiger, among others. A number of economic geographers and city planners have examined

related issues. Once again, this is progress, but many issues are unresolved.

- When the broadband debate heated up, policy advocates were "shocked" to learn that they knew so little. Yet, this had been the norm for some time. Does the availability of broadband encourage entry of businesses into a local area? How does the provision of multiple suppliers affect pricing and productivity of local firms? Do these investments help some types of industries grow, and, if so, which ones? In general, we did not know the answer to such questions a few years ago and we still do not know.

- It is an open question whether regional difference in using the Internet at home affects local economic conditions. Most researchers believe it does, but that is hard to measure in practice. Here is an illustration as to why: A recent Northwestern dissertation examined whether differences in the diffusion of the Internet altered the prices paid for airline tickets from cities. As it turned out, the answer is a long story. Yes, the Internet did make a difference, mostly at the mid and upper tail of the distribution of prices paid, but not everywhere on every route. However, to see it more clearly, one had to control for the "Southwest" effect and the "hub home" effect, and so on.
    - Other recent work has started to see if there are price effects in other areas of the economy where we might expect to find it, such as bookstores, travel agencies, newspapers, and so on. Once again, there is much to do. There are a number of data sets about this topic, which I will describe below.

Even after a brief overview it is clear the situation is not completely void of data. Readers should not view this area as a vast wasteland. Rather, it is an area of study where data typically support an imprecise impression. It is the type of environment in which a lazy journalist will give up quickly, and a good journalist will find considerable room for speculation. More to the point, policy analysts and researchers will find lots of frustration

because the data usually cannot support inferential or predictive models with any precision, so it becomes difficult to guide policy choices.

There is variation around this trend, of course. A few areas of study are exceptional and a few are not. To appreciate that variation one should be more precise about what is there. I will begin with macro-economic data, then go to more micro-policy questions.

- The Bureau of Economic Analysis has tried for some time to improve the situation. For example, they now have some estimates of the differences in IT investment across industries. This is the data largely found in Dale Jorgensen and Kevin Stiroh's (2006) extensive analysis of the contribution of IT to the US economy. BEA also has tried to develop serious estimates about differences in labor markets for programmers across the US. Both of these are recent improvements.
    - However, both of these datasets measure investment in IT at a broad level. They do not specify anything special about the Internet, or technical advances in any of the sub-components of ICT. This precision about the statistics shapes the interpretation. This data is useful for learning about whether investment in ICT matters, not about whether technical change in ICT produces growth or productivity advance. For the latter, one would need to collect statistics that distinguish between a PC, a router, software, type of software, and so on.
        - It should be noted that these data at BEA are a step up from the past. They come from programs initiated in cooperation with other federal agencies, such as the US Census and the Bureau of Labor Statistics. They require planning, foresight, employee effort, and especially, funds for conducting surveys. All these parties deserve credit for coordinating prior initiatives. These are inherently big projects.
        - I do not want to gloss over the challenges about whether this data is accurate. The measurement issues here require careful treatment of many difficult topics. First and foremost, economic research

must drop its excessive focus on counting hardware as the primary source of value. Much of the value in a device resides in its software. Second, much of the expense in a business resides in maintaining the hardware and software, but that frames a question: how should such expense go in the national accounts? It is just like an investment that seeks to prevent depreciation of assets due to deterioration, but it seems strange to treat human labor a capital expense instead of a variable cost of production. Third, the valuation questions are hard. How does one value custom software? What is the deflator? The market value for software does not need to reflect its service value if the market price declines due to technological obsolescence. At what rate does the value of packaged software decline if its service flow does not decline?

- Just recently these initiatives led to the first survey of software at firms, (at the level of the firm, not the level of the establishment). This was motivated by a desire to bring software into the estimates of capital accounts in the US. This project is well worth doing. It is not exactly like trying to find all the dark matter in the universe, but it has many similarities. For years everyone knew software was there and it was big and valuable, but there is no precision beyond those generalities.
    - It turns out that the best estimates were large, in the tens (hundreds?) of billions of dollars.
    - It turns out that figuring this out for just one year, 2003, was almost impossible, even using the best statistical agencies in the world, the US Census. Even the practical issues are mind-boggling. For example, who is the right person to ask in a company when there is no CIO, and how should the US Census phrase their questions? Many multi-establishment companies have only the vaguest estimates of how much software they own and operate across all their establishments, so how much faith should the Census put in a survey answer? Given the uncertainty for specific

firms, how should the Census weight the results to make estimates for the entire economy?

- As a result, figuring out differences between industries, for example, has turned out to be near impossible.
- Census is trying to improve what they do, but nobody has any ambitions for developing a precise estimate for the growth in software nationwide, one year to the next. Maybe we will get one of changes every few years.

- Overall, the state of economic knowledge about the stock of software in the US is a lot like the experts declaring that the universe is expanding, but they are unsure about the rate of expansion. Yes, the universe could be expanding a little or a lot, but there is no way to know yet.

This state of things has implications for some of the topics Internet researchers care passionately about. For example, analysis of the productivity improvement from investment in ICTs will be hounded by questions about the extent of mismeasurement at the foundations of the exercise. And the answers will go to the heart of analysis looking at the timing of economic change, whether investment yields economic benefits today or tomorrow, whether the macro-economy shapes investments today, and so on. If anybody thinks this type of question is trivial, please recall that Alan Greenspan wanted to know in his own day whether investments in ICTs would yield high returns and whether the returns were measured well or not. The answer shaped whether he tried to keep interest rates low enough to encourage such investment. Greenspan decided the returns were high and under-measured, and for numerous reasons kept interest rates low – i.e., during the dot-com bubble, then started to raise them after December, 1999, as it began to burst. If advances in ICTs continue to emerge (as most of us expect), then the same questions will continue to dog future chairs of the Fed.

Related, in the present era there is a call to learn about the growth in availability broadband, as well as its use and quality. This is probably a feasible project for a US federal agency with experience conducting large surveys, like the US Census. However, to be fair, the “dark matter” issues with software are a bigger priority because, quite

frankly, it involves more money and has been mismeasured for quite a long time. Also, in a time of limited budgets to collect data, it is quite understandable why that remains the priority.

Which motivates a related question: Until such time as budgets free up, what data can researchers use?

- Outside of the surveys done by CRITO, mentioned above, the best available data about business use of ICTs is the data provided by Harte-Hanks, which conducts a private survey of ICT use by business. Several groups of researchers who have used this survey, as noted earlier.
    - To be fair, this data is reasonably good at measuring big discreet things, such as whether a business establishment uses the Internet (though, it is not particularly good at measuring "how much").
    - This survey was NOT designed for academic researchers doing economic analysis. It is designed for marketing purposes. Its quality for marketing is fine, but for our purposes the quality of the survey is deteriorating, especially recently, because certain data are no longer collected. It is hard to say that with any anger, since economic researchers are rather small purchasers of these data. Why should a private marketing company care about public policy uses of its data?

Finally, I would be remiss if I did not mention two initiatives at the US Census.

- First, there is a special survey conducted by the US Census about electronic commerce use at business establishments. It was a supplemental survey in 1999 of networking use by 10% of US manufacturing establishments. An employee of the census, B.K. Atrostic & many of her colleagues at the Center for Economic Studies, have done some excellent analysis of that data.
    - However, this survey was not economy wide and it was not followed with additional surveys because of budgetary pressures. So the situation is far from ideal. For example, we do not know anything comparable about what has happened to the finance sector of the US economy, which is among the biggest users of

advanced IT in the US economy. We also do not know much about IT use in the health care sector, the services sector (such as lawyers, accountants and consultants), education sector, warehousing, transportation services, and government agencies. That is also a lot we do not know. (By the way, this is not asking too much. Remember, by comparison, *we do know* who uses cement and concrete and how much they use each year).

- This survey will become rather obsolete for contemporary policy formulation soon. This should concern many people because there has not been any plan put in place to follow it up and find out how much things have changed in ten years, even in manufacturing.

- Second, Census has undertaken regular estimates of "electronic commerce" across different industries. These estimates began in 2000 and have continued since. This project is known as E-Stats. See http://www.census.gov/eos/www/ebusiness614.htm. It provides estimates of electronic shipments and revenues for many four digit industries. It is an admirable initiative, as with the others noted earlier. It has improved the situation.
    - At the same time, the numbers are limited. They are large enough to suggest that the amount of dark matter must be large. They also suggest that big changes are afoot in some industries and not others. It is less clear how to relate these data on electronic commerce to changes in productivity and concerns about economic growth. As far as I know, so far nobody has tried to look for links.

One of the more remarkable statistical accomplishments has been in the areas concerning the availability and use of the Internet by households. To be fair, our understanding of households is better than our understanding of business. We have good statistics about household use.

- Probably the most widely used statistic is one compiled by the BLS about the adoption of the Internet at households. Collection of this statistic started in 1997,

as an additional question to a supplement to a regular survey of households. This supplement had been devoted to finding out about PC use at households in 1995, so it was naturally and easily retargeted at Internet use. The basic results have been widely reported in a series of reports published by the NTIA in the Department of Commerce. At some point virtually every researcher in this area has used this data and for a wide variety of purposes.

- This series was the best continuously collected historical statistics about changes over time in the US. It uses BLS sampling procedures, which have been refined for years and, as a result, everyone trusts that they are representative samples of the US – which is not a trivial feature, because it allows one to weight results for "national averages" without major concerns about biases. The sample is also large enough, so the underlying data give large "cells" for the two hundred largest metro areas in the US.
- To be fair, these surveys had issues too. They do not contain very good data about prices or expenditures or about service quality. Many questions were discontinued. The survey also asks only basic questions about what users do on-line. Sociologists of the on-line world find the data completely uninformative about the issues they consider central.
- This triumph comes tinged with recent absurdity. This survey continued until a few years ago, though it is my understanding that there is no plan to continue it further. To most of us in this area this decision looks myopic to the point of tragic.
- Tragedy was partially averted by charity. There is now another survey under way, done by the Pew Charitable Trust, about household use of the Internet. It asks extensive questions, even more than the government survey, about what users do with their Internet, so it is quite informative about the changing nature of the on-line experience for Americans in the new millennium. John Horrigan, the principal researcher in charge, has been willing to let others look at this data, and has initiated several projects, including with Ken Flamm, among others.

- What else can someone do? A very small set of researchers have also used household surveys conducted by Forester, a consulting company. This includes Austan Goolsbee, as noted above. Researchers such as Goolsbee and Klenow (2002) and Sinai and Waldfogel (2005) have made progress. A recent student at Northwestern, Jeff Prince, now faculty at Cornell, also used this data for his dissertation on PC purchasing. Occasionally researchers have been able to get reasonable data of surfing behavior from private firms such as Media Matrix. Both Goldfarb and Hitt, for example, are among those who have done this.

- Related to this last set of comments, there is a small and growing literature among academics about the interplay between electronic retailing and brick-and-mortar retailing, occasionally with household behavior thrown in, often with a marketing question motivating the investigation. Virtually all that data come from non-government data collectors, or directly from companies. There is also a larger and still growing academic literature on Internet auctions, mostly about E-Bay. Virtually all that data come from spiders. It would take many pages to summarize this literature, which is a good thing. To my knowledge, there has not been much connection between this literature and government collected data about economy wide trends in electronic commerce. There also has not been an attempt to connect these papers back to the fundamental concerns about productivity.

When it comes to describing the availability of the Internet – in other words, who offers what service for what price and in what locations – the situation is mediocre. During the dial-up era a quirk in the telephone system helped researchers, but during the broadband era the situation has gotten worse.

- The best data on the availability of dial-up Internet access in the US was done by Tom Downes and myself during the period 1996 to 1998 (Downes and Greenstein, 2002, 2006). We were able to figure out the location of every commercial Internet provider in the US because the area code and prefix for a firm had an association with a unique telephone switch in a unique local calling area.

- Frankly, it was a real thrill to have one of our maps published in Time Magazine, but we also got tired of spending six months of the year (and a full time RA) doing this one activity, which is what it took to make those maps. The last map involved over 65,000 phone numbers. So we stopped after the fall of 1998.
- To my knowledge, nobody has bothered to try again except in a few special places like the Appalachians and the Mississippi delta or West Texas, where the policy issues are especially difficult. Sharon Strover and colleagues have mostly done this work. It is also my impression that some state development agencies know about their own situation, but there is no general sharing of this information across agencies. This is still a relevant question since broadband has failed to deploy in many low density areas, but nobody knows how good/band is the quality of access in such areas, especially for alternatives such as upgraded cell phones.

- The FCC publishes data on the availability of broadband across the country (see http://www.fcc.gov/wcb/iatd/comp.html). It ostensibly measures how many providers offer broadband in specific zip codes. It recently began to distinguish between use at households and business, and it does give state statistics. However, it does not say anything about the geographic scope of availability within a zip code, nor does it say much about the size of the firms making these offers, nor about their prices. It recently added some coarse information about bandwidth, which is a mild improvement.
    - Of course, bandwidth is not the same as speed at the end point. How fast a response does a gamer get to ping he/she sends out to another player in another part of the country? That depends on the bandwidth of the connection to the end-point, to be sure, but it also depends on the quality of the connection, routing, and so on, and it will also vary at different times of the day. When engineers

complain that they know little about the quality of the network in the US, this is part what they mean.

- As it turns out, according to the footnotes of Ken Flamm's (2006) most recent paper using the data about availability, this data does not particularly measure availability well without a lot of effort by a researcher, which means the published data are historically inaccurate about the details of how broadband diffused.

- The deployment of wi-fi has been proceeding apace for some years now. There are on-line sites listing which cities have deployed city-sponsored networks. There are also a few studies of the deployment in some neighborhoods in Los Angeles and Chicago and a few other cities (basically showing that richer neighborhoods have more). National statistical agencies have not followed this in detail. Christian Sandvig is one of the leading researchers in that area. We know some basic things, but we do not know about the patterns of deployment across the country. The amount of dark matter has to be large because the equipment providers tell us they are doing a good business, but beyond that it is hard to tell what has happened.

The pricing of Internet access: an illustration

Let me end with a small illustration of the state of things. I will use the price of internet access as the example. I want to do this because this aforementioned list of issues is long and somewhat general. I thought a very specific example might help illustrate the broader issues in play.

No statistic would seem to be as important as knowing the price of the Internet. After all, prices are the bread and butter of economics, the primary signal that markets have for indicating scarcity. You might reasonably respond: how can the Internet have a price attached to it? As it turns out, there is such a price. It is called the price index for Internet access.

The price of Internet access attracts interest for two reasons. First, the value of the Internet is unknown. The access price is one place where most users give up money in exchange for free Internet services. During the first decade of the commercial Internet the typical household spends more than three quarters of the time on line at free or advertising supported sites.[1] Second, most of the value from the Internet is embodied in this index. Other than through access fees, there is little data from which to infer user's valuation of these sites. This component of the web – that is, access fees – also produced the highest fraction of total market value in the first decade of the commercial Internet. Though subscription based services are growing, the revenues spent on access fees swamp subscription services in magnitude. Fees for advertising are also growing and may exceed access revenue soon, but for the first decade they did not.[2]

The official price index for Internet access in the CPI started in late 1997, reasonably early in the market's growth (an initiative for which the BLS should get credit). In 1998 the access market was a 10.8 billion dollar industry, with roughly half that revenue coming from access fees. As of 2004 it is a 24 Billion dollar industry.[3] In other words, it is a growing industry. Once again, that growth highlights the importance of deflating revenues properly. Is the true growth much large or much smaller than the changes in nominal growth of revenues?

Table 1 contains a summary of prices for Internet access in the United States.[4] This is the number one would use to deflate the growth in revenue over time to figure out how much of the growth was real and nominal, with the latter being denominated in

[1] Goldfarb, 2002, contains click stream data verifying this trend in detail across many categories of uses. It is also commonly verified by observers web use, such as Media Matrix. Also see O'Donnell for an estimate of the relative magnitude of the value chain behind access and advertising on the Internet.

[2] According to NAICS revenue for 514191 and 514199, access fees exceed advertising by three or four times. For example, total advertising for 1999 was 1,355 + 1,477 = 2,832, as compared with 8,979 under NAICS 514191 alone. See the 2001 statistical abstract, table no 1151, Census Bureau.

[3] For 1998 see revenue for NAICS 514191. See the 2001 statistical abstract, table no 1151, Census Bureau. This lists revenue of $10,882 million for 1998 and $18,025 for 1999, with access fees accounting for $5,499 and $8,979 respectively. These revenue estimates are based on the 1999 Service Annual Survey, Information Sector Services. In the 2004 Service Annual Survey (released by the U.S. Census Bureau on December 29, 2005) NAICS 514191, Internet Service Providers had $10.5 billion in access fees in 2004; NAICS 5132, Cable Network and Program Distribution, had $8.6 billion in Internet access service; and NAICS 5133, Telecommunications Carriers, had $4.3 billion in Internet access services. This does not count the revenue for other on-line services.

[4] This is "computer information processing services" at http://data.bls.gov, or see "On-line information services," NAICS 514191.

constant dollars, as any Econ 1 textbook says we ought to do. I will not provide a full assessment of this table, but I do not want to indicate the issues with it.

Table 1. US Internet access price index.[5]

YEAR	12/97	12/98	12/99	12/00	12/01	12/02	12/03	12/04	12/05
INDEX	100	103.3	96.0	95.7	100.3	99.6	97.6	97.2	94.2

Table 1 says that the price of the Internet has declined by less than 6% in 8 years. That is not a misprint. At the same time, this is not the price movement usually associated with the diffusion of a revolutionary technology. More to the point, this index seems almost unrelated to the improved experience of virtually every online surfer.

You might be concerned that this index has some error in it. To sharpen the focus on the right issue, consider several other closely related categories of good from the same BLS sources. During the same time period the official price indices for the US had the following patterns: computer software and accessories went down 42 percent; personal computers and peripheral equipment (which does adjust for quality) went down 88 percent; telephone hardware and calculators and related consumer items went down 55 percent; and wireless telephone services went down 35 percent. These patterns would lead one to conclude that there was lots of price change in everything except the Internet. How can it be that Internet access did not get cheaper, but cell phones did, phone equipment did, PCs did, and software did?

First, let's be up front about the accuracy of the data: This data is collected with exactly the same procedures as those used to collect the other indices. On one basic level it does exactly what it is supposed to do. It is procedurally correct.

On the other hand, it includes some measurement error. To put it blithely, it does not adjust for quality. Some of this is straightforward to recognize though tricky to fix, such as adjusting for upgrades in bandwidth, as from 28k to 56k, (which largely occurred in 1998 and 1999), or from 56k to broadband speeds of varying qualities and availability (which has been occurring since about 2001). Some of it is quite subtle and difficult to

[5] This is the monthly price quote, as indicated, under "Internet Services and electronic information providers." See http://data.bls.gov/.

fix. For example, how should the BLS measure quality when new broadband comes on line, and half the country switches to this new source in a short period? Some issues are probably impossible to solve. For example, the index does not adjust for the quality of the many free compliments over this time period, such as in the search engine Google or the portal Yahoo!.

These errors are not the fault of the person who collects and compiles the data. He is doing exactly what he is supposed to do, making a time series for the Internet that is comparable to any other price series, using the data collected by standard BLS procedures (which weight highly the prices of large and stable providers, such as AOL). The fault lies with the lack of funds to initiate anything else, and with a very conservative philosophy throughout statistical agencies in the US, who do not want to be criticized by overseers for even a slight deviation in procedures (or industry associations – yes, this does happen with alarming frequency).

While that may seem like a small thing, this little error shows up in all sorts of ways. For example, because the rate of growth in the revenues do not appear particularly high (because they are not properly deflated for improvement in quality) I have heard some macroeconomists maintain that the contribution of the Internet to the economy was small or, if large, short-lived. That conclusion looks right in a literal way, based on the data available, unless one recognizes the error with its collection. And that is my point: Unless a researcher is a super expert, it is hard to recognize. Should all Internet economists ask all the macroeconomic experts in the country to be experts on the mis-measurement of broadband in the official US price index for the Internet? Of course not. We should fix it so they reach the right conclusion when they read the numbers literally.

Think about another historical implications of that simple observation: it is one thing to say the Internet bubble was a short-lived phenomenon equal to the tulip craze in Holland centuries ago; it is quite another to say that the Internet bubble was similar to other halting ups and downs that accompanied the building of the railroads. In the first case, little economic value was created. In the latter, enormous value was created, albeit not in a calm fashion. Right now, a literal reading of the data favors the former interpretation. Most Internet insiders think that is wrong, but we do not have much data to support our case.

As another example of the difficulties with measuring prices, recently the OECD tried to issue a comparison of prices for broadband across different countries, which is a wonderfully ambitious and useful thing to do. Of course, the exercise is full of footnotes and qualifications, not only because the researchers were careful, but for a simple reason: there was not comparable data available across countries for even the most basic service, the price of access. Imagine how much more useful this would have been if such data were abundant?

My broad point is that something as basic as the price index for Internet access is far from perfect, but it is extremely useful. More broadly, something as basic as this has not standardized the basic definitions for price, quality and its changes over time. Is it any surprise that so much else has been neglected as well?

As one epilogue, I might add this: the recent announcements by AOL about change to its pricing – it is moving to advertising supported service in response to losing customers to broadband – will bring this index down dramatically for the first time in a long time. As of this writing, the index fell from 93.1 in September of 2007 to 77.2 in December, 07. This is the most dramatic thing that has happened to these data over several months since the summer of 1999, when AOL attempted to give price breaks to former CompuServe users after AOL merged with CompuServe.

It is a good thing if this index falls. However, the timing is all wrong. It excessively weights the behavior of the large and established firms, such as AOL, missing the other dramatic things that have already occurred, such as half the country switching to broadband. It has delayed recording the timing of a change, putting the dramatic fall in prices many years too late.

Unanswered questions

One might reasonably ask whether any of this absence of data matters. After all, maybe it is quite ok that we know more about cement and concrete than ICT in a world where we only have so much time and money to spend on collecting statistics. Maybe the basic building block for highways is more important for the economy than the basic

building block for information super highways. In response, let me give a little personal anecdote and generalize from it.

The OECD occasionally publishes a statistic comparing Internet household adoption across many countries. The last time this occurred it sent a few US Congressmen into panic because Korea, Iceland and a dozen other countries came out with larger numbers than the US. As a result, a few news reporters found the topic interesting and then searched around to call someone. Occasionally, I fielded that phone call.

The very first thing to say to a reporter is this: cross country differences in household adoption of broadband tells us about the marginal adopter and that is about it. That is useful, but not the end of the story. It is not especially informative about use by the intensive user, which are the users who generate most of the traffic. It also does not tell us about business adoption and use of broadband, which is where the productivity gains come from. Moreover, it is uninformative about the productivity impact of the Internet, because, once again, the biggest impact to productivity will come from the large adopters. I could go on, but the basic conclusion is straightforward. This index is but one piece of data. It is good to have, but it is not very informative about all of the important questions of our time.

Suffice to say, most major US newspapers have not run headlines declaring "OECD statistic is helpful but ambiguous" with a sub-headline "Congressmen agree to stop panicking and ask for better data." Usually the reporter hangs up and calls someone else who is willing to be a bit more pithy about this topic.

It is just a simple statistic, but it illustrates a broad point. Simple statistics can shape policy, especially by Congressional staffers who are not going to spend much time studying the issues. If the statistics are shallow, the policies will be too.

So let's talk for just a minute about how little we know about US broadband. The best statistics on broadband come from the FCC report on competition, available at http://www.fcc.gov/wcb/iatd/comp.html in a report for high speed services for Internet access. It is something, but it is also unsatisfying in many ways. It tells us something about the number of lines available, the number of providers available, and the type of adoption patterns found across states, and some basic distinctions between the market for

business and households. It uses a comparatively low bar for "high-speed", and recently added some gradations for that.

Here, for example, is a short list of rather basic issues in the broadband policy for which we do not have even the most basic data:

- What industries make the greatest use of broadband? How has this improved their productivity, if at all?
- Which types of firms benefit the most when broadband becomes available? When broadband prices decline or quality improves?
- What areas of the country benefited the most from broadband in the last decade? Where have local taxes improved the most as a result?
- Which areas were hurt the most from broadband's diffusion (e.g., central Nebraska lost a large part of its telemarketing industry)?
- While the data tells us a bit about different types of providers (e.g., DSL versus Cable), it does not tell us much about their average quality, even though it is well known that cable firms have improved their speed over time. What is the average price per speed obtained by the typical user?
- What technologies generally go in hand with broadband? Which technologies enable or nurture a cluster of other technologies? What other investments follow?
- What happens to local wages for local progammers as broadband diffuses? Do they change at all? How about earnings at other technically-intensive industries?
- How has business use of broadband shifted in response to concerns over information security? Has this varied by industry?

Summary

The state of data collection for the Internet economy is mediocre. I have tried to indicate that this is pervasive throughout many of the key questions in Internet economic policy. I have tried to show that some of the data used today by researchers comes from the US government, but quite a lot also comes from private sources. Moreover, some of the most useful initiatives undertaken by US statistical agencies are difficult, in need of money and effort.

More to the point, US policy formulation across a range of issues is poorer as a result. This gap in knowledge pervades many issues that come up at the FCC, the US Congress, the Department of Commerce, the Bureau of Economic Analysis, the Internal Revenue Service the Bureau of Labor Statistics, the National Science Foundation, and elsewhere.

In closing I want to remind the reader about the purpose of this essay. I wanted to describe the range of places that lack statistics. I wanted to identify places where we were far from any reasonable ideal. To illustrate that point I have tried to be an equal opportunity offender, citing and critiquing statistics coming from a range of government sources, in this case, BLS, the Census, the NTIA, and the FCC. I was not trying to single out these agencies for error or mistake, but to show that they are trying to address challenging issues with meager resources. These illustrations were at hand and seemed like enough to get the point across. More to the point, these agencies need help. I hope it will motivate others to initiate projects that will fill gaps.

Further reading from literature reviews:

Chris Forman and Avi Goldfarb (2006), "Diffusion of Information and Communication Technologies to Business," in (ed) Terrence Hendershott, *Handbook of Information Systems, Economics and Information Systems, Volume 1*, Elsevier.

Zvi Griliches (1994), "Productivity, R&D, and the Data Constraint," in *American Economic Review,* 84, pp 1-23.

Shane Greenstein (2005), "The Economic Geography of Internet Infrastructure in the United States," in (eds) Martin Cave, Sumit Majumdar, and Ingo Vogelsang, *Handbook of Telecommunication Economics, Volume II.* Elsevier.

Shane Greenstein and Jeff Prince (2007), *Forthcoming* (and Jeff Prince), "The Diffusion of the Internet and the Geography of the Digital Divide," in (Eds. Robin Mansell, Danny Quah, and Roger Silverstone), *Oxford Handbook on ICTs*, Oxford University Press.