OECD publishing

Please cite this paper as:

Lehr, W. (2012-04-13), "Measuring the Internet: The Data Challenge", *OECD Digital Economy Papers*, No. 194, OECD Publishing, Paris.
http://dx.doi.org/10.1787/5k9bhk5fzvzx-en

OECD Digital Economy Papers No. 194

Measuring the Internet

THE DATA CHALLENGE

William Lehr

OECD DIGITAL ECONOMY PAPERS

The OECD's Directorate for Science, Technology and Industry (STI) undertakes a wide range of activities to better understand how information and communication technologies (ICTs) contribute to sustainable economic growth, social well-being and the overall shift toward knowledge-based societies.

The OECD Digital Economy Papers series covers a broad range of ICT-related issues, both technical and analytical in nature, and makes selected studies available to a wider readership. It includes *policy reports*, which are officially declassified by an OECD committee, and occasionally *working papers*, which are meant to share early knowledge and elicit feedback. This document is a working paper. The opinions expressed in this paper are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

STI also publishes the OECD Science, Technology and Industry Working Paper series, which covers a broad range of themes related to OECD's research and policy work on knowledge-based sources of economic and social growth and, more specifically, on the translation of science and technology into innovation.

OECD Digital Economy Papers and
STI Working Papers are available at:
www.oecd.org/sti/working-papers

MEASURING THE INTERNET: THE DATA CHALLENGE

William Lehr*

Abstract

This paper reviews a number of the challenges and opportunities confronting analysts interested in measuring the Internet and its economic and social impacts. It identifies several additional challenges to the measurement issue, in addition to all of the normal problems one expects when measuring information and communication technologies (ICTs). These challenges are related to: *(i)* the rapidly changing nature of the Internet, *(ii)* the need for more granular data in order to understand the complex nature of the Internet, and *(iii)* the phenomenon of big data and the resulting ability to measure almost anything.

* Massachusetts Institute of Technology, wlehr@mit.edu. The views expressed in this paper are those of the author and should not be attributed to the OECD or its member countries.

TABLE OF CONTENTS

MEASURING THE INTERNET: THE DATA CHALLENGE

Measuring the Internet and its economic and social impacts presents a number of significant data challenges. These include all of the same ones that have bedeviled efforts to estimate the economic impacts of computers, broadband, and other information and communications technology (ICT) components. Solow's 1987 paradox still applies.[1] Although a number of firm-level studies were able to demonstrate the positive productivity contributions of computers in the 1990s (Brynjolfsson and Hitt, 1998; Lehr and Lichtenberg, 1999,) it was not until after 2000 that the large contribution of ICTs to economic growth was demonstrable in macroeconomic data (Oliner and Sichel, 2000; Jorgenson, 2001.) This earlier experience offers several important lessons that are worth remembering as we turn our focus toward measuring the Internet. First, we will likely only be able to reliably estimate the economic impacts *after* the fact. Businesses and government policymakers needed to invest in ICTs *before* economists were able to demonstrate that economic impacts were, in fact, positive. Second, the first and best evidence of economic impacts is likely to come from micro-data (firm or smaller) before it shows up in macro-data (industry or national).

The changing Internet

To measure its economic impacts, we must first measure the Internet, which is problematic because the Internet is itself changing. To date, the focus has been on measuring the availability and adoption of Internet access, first dial-up and now broadband. Heterogeneity in adoption behaviour across firms, households, and industries provided a reasonable proxy for use, which is what is ultimately of interest, since economic impacts arise only consequent to Internet usage.[2] As the Internet evolves to become basic infrastructure and adoption saturates, the Internet economy will become increasingly indistinguishable from the overall economy. What will matter is how different firms, workers, or consumers utilise the Internet and measuring that is inherently more difficult. We will need more granular data, at the business location or unit level, or even better, at the worker activity level to understand how the Internet is being used in productive activities. This is due in part to the fact that the Internet enables the creation of virtual organisations and flexible outsourcing of business activities, blurring the boundaries between firms and markets, between work and social life.[3]

The growth of the self-service economy and the changing role for consumers as producers of media content, participants in product design, promotion, and transaction processing illustrate this phenomenon. Separating consumption from production when the boundary between firm and customer blurs requires ever-more detailed information about the specific activities being undertaken. Increasingly detailed business and labour surveys will be needed to track how the Internet is being used to accomplish the varied tasks that go into business production. Initially, we may focus on the time spent using the Internet in different business functions (*e.g.*, research and development, supply chain management, retailing, or G&A) or worker activities (*e.g.*, web browsing, word processing or communications).

However, in our increasingly *always-on/everywhere connected* world, we may find the time not spent on the Internet the exception. We will need even more granular data on the location and intensity of usage – the volume of traffic, input/output activity, and perhaps even, the level of attention of the worker (*i.e.* was use of the Internet intrinsic to performing the activity or in the background?).

Moreover, as we seek to measure the Internet itself by counting first the number of broadband connections, and more recently, the number of broadband connections by speed tier and for increasingly smaller geo-locations, we will find that the resources we need to track are becoming more varied and complex.

For example, consider some of the more important trends in the Internet ecosystem. These include: mobility, cloud computing, social networking, and sensor-nets. All of these are central to current strategic business decision-making about ICT usage and to enabling the heralded future of *smart everything* (grids, homes and business processes; or, energy, healthcare, transport and government), but none of these are readily measurable via our traditional metrics focusing on data like line counts, fibre-miles, megabytes of traffic, or IP addresses. Furthermore, the entities we will need to sample to collect information are changing also. Internet Service Providers (ISPs) are becoming more heterogeneous and the value chain for the Internet grows more complex. Increasingly, non-communications-sector participants may be responsible for key decisions about the deployment of new Internet infrastructures. This may include energy and transport providers deploying smart grids; healthcare providers deploying eHealth data systems; or resource companies deploying resource management systems.

Fortunately, with the growing recognition that the Internet constitutes basic infrastructure, tracking the state of critical infrastructure components, their availability, costs, investment, and usage will be motivated by more than a desire for better policymaking for economic development in the information society. Communications regulators and service providers will need to collect and track such data, constructing detailed and very granular Geographic Information Systems (GIS) similar to what currently exist for other key infrastructures like electric power distribution grids, water systems, and roads. They will need such data to manage network performance and infrastructure investments. ISPs and other value-chain participants such as Akamai, Amazon, Google and Netflix are all instrumenting their networks to provide detailed real-time data on a growing number of traffic attributes to manage individual and aggregate traffic flows across time and across the end-to-end Internet.

Although significant volumes of data are being collected using tools like the Deep Packet Inspection (DPI) products from companies like Sandvine[4] and Arbor Networks,[5] there are no generally agreed standards on metrics for measuring or classifying traffic, and the appropriate business uses for such data.[6] Even seemingly simple measurement questions such as “how best to characterise the 'actual' speed of broadband access services?” prove to be quite complex (Bauer, Clark and Lehr, 2010.) In response to the inadequacy of previously available tools, regulatory authorities first in the United Kingdom, then in the United States, and now in the European Union are deploying traffic measurement infrastructure from SamKnows.[7] When such platforms are extended to mobile devices, the potential for finer-grained traffic measurements and characterisation will expand significantly. As another example, consider a project in which public transport buses in the Madison, Wisconsin area have been equipped with mobile wireless broadband access, as a benefit to riders (Sen *et al.*, 2011.) This platform also is being used to make real-time network measurements for multiple mobile service providers in the area. Such information and platforms can be used to obtain a very granular picture of mobile broadband performance by time of day and location throughout the area.

Difficult questions include: how much information is needed for different purposes (sampling)? and how long is it useful to keep the data (what sorts of trend analysis is worth doing)? ISPs maintain very detailed data for relatively short periods of time in support of their business operations, but most compute summary statistics and traffic aggregates for longer-term data storage. Whether the loss of the historical granular data matters or not depends on the questions you want to ask. It certainly poses a challenge for *ex post* forensic analysis of network behaviours that might be of policy interest for regulatory enforcement.[8]

A still more difficult question, which we will return to later, is who should have access to the granular information collected. For example, few see a problem with making data publicly available on which ISPs provide broadband access services aggregated over relatively large geographic regions (*e.g.*, provinces or countries); however, many might disapprove of posting detailed maps on the Internet that show where key network components are located. Such information might be used for terrorist attacks or to target competitive investments.

Finally, the significant trends changing the Internet will have profound implications for Internet measurement and metrics. For example, consider the challenge of comparing the performance of fixed and mobile broadband services: the latter may vary in geo-space as well as across time (Lehr, *et al.*, 2011.) Mobility allows us to consider more localised contextual information (the micro climate) that may have bearing on our decision-making (*e.g.*, do I need an umbrella here at this particular time and place and given what I am doing?). Just as the mobile telephone personalised telephony, mobile broadband has the potential to personalise/individualise Internet usage.[9] When augmented with sensors, cloud-based resources, and the other components that support pervasive computing environments, it becomes feasible to undertake real-time, dynamic, interactive control/optimisations (of traffic on a highway, of energy usage in a household, or of pricing in a market) to customise system performance to local conditions in time, space, or in response to other contextual factors.

Data, data everywhere

The need for more granular data for meaningful insights into an increasingly pervasive Internet will be answered, in part, by the Internet itself, which will be an ever-more powerful platform for collecting, processing, managing, and presenting all kinds of data – not just data about the Internet. The growth in deployments of sensors and mobile platforms, cloud-based storage and processing, and social-network-fueled crowd-sourcing will make it possible to collect massive amounts of data on almost everything. The data will be collected automatically by passive and active sensors –monitoring everything from road traffic to forest growth, from web-browsing clickstream data to home appliance usage; and by armies of individual users blogging, using smartphone measurement applications, and completing web-based surveys and tests.

The Internet is a powerful platform for processing and managing the data. Web-crawling software can extract and summarise huge volumes of data automatically and at relatively low cost. For example, consider the "WeFeelFine" project that has been culling blog postings for sentences that contain the phrase "I feel" or "I am feeling" since 2005.[10] Whether one views the project as a work of art or a database of human feelings, the project is illustrative of the future in several respects. It demonstrates the viability of automated, large-scale data collection and processing capabilities. The code for the system is open-source, available for use under a Creative Commons license, to be used and potentially modified for other data collection and presentation projects. The system automatically correlates the blog postings with the location of the author and the weather at the time, allowing one to see whether the weather, country-of-origin, or age of the authors has an impact on the feelings expressed. While one may quibble about the accuracy of any of this data or the inferences that might be drawn, the basic functionality for data collection and processing suggests what can be accomplished relatively inexpensively. Finally, the authors offer a number of data visualisation tools that illustrate the Internet's potential for novel ways to present and interact with the data. The data may be viewed as a scatter plot of moving coloured dots that one may click on to reveal the expression of feeling from the blog posting and potentially the picture that accompanied it, or as a collage of pictures from multiple authors associated with a particular feeling, or as various charts or graphs.

As another demonstration of the need and uses for new visualisation tools, consider the work of Drew Conway who wrote a program to analyze the WikiLeaks documents for mentions of attacks in Afghanistan by year and then plotted those on a map of Afghanistan (Schactman, 2010). This approach highlighted features of the data which would have been more difficult to make sense of otherwise, and of course, had an emotional appeal that is much more immediate than a list of dates and places might have been (a point I shall return to further below).

Finally, consider the work of Rosalind Picard and her colleagues on using webcams to identify facial expressions and measure heart rates. The quality of the cameras has progressed to such a point where such low-cost remote sensing strategies are becoming practical. Such tools can be used for real-time health monitoring[11] or marketing research.[12] The potential for automating data collection and analysis – for automating research – offered by such tools is great but also raises important challenges.

Big data, new problems

The ability to measure almost anything, anywhere and in real-time is giving rise to massive new collections of data. Noteworthy examples include the clickstream and social-networking transactional data being generated continuously and globally. The data sets can be massive. For example, one source reported that Walmart's transaction databases exceeded 2.5 petabytes (10^{15} bytes) and that Facebook included over 40 billion photos.[13] When data sets get that large, they stress the limits of our computer hardware and software resources. Working with such large collections of data requires new tools such as cloud-based storage and parallel computing frameworks.[14] When data gets sufficiently plentiful, data mining may replace traditional scientific methods.[15]

In addition to the computational challenges, there are significant skills deficits when it comes to analyzing and presenting the data from such large data sets. For example, when adding up a large sequence of numbers, it is important to consider the round-off error that arises as a consequence of the floating-point representation of real numbers used by digital computers (Judd, 1998.) Empirical researchers unaccustomed to working with big data sets are often unaware of such problems. Likewise, revealing structure in the complex data often requires recourse to new dynamic graphing and visualisation techniques that may be completely novel to many empirical researchers (*e.g.*, when working with mixed geo-spatial/temporal data).[16]

The availability of data will drive policymakers and researchers to expand the range of questions they may seek to answer. Economies and markets are complex systems, but traditional economic methods have focused on simplified models based on limited data. The data limitations were often imposed by cost and observability considerations. The result is that traditional economic methods often do a poor job at explaining the dynamic behaviour of complex systems (Tesfatsion and Judd, 2006.)

In the context of growth economics, this is motivating a new breed of researchers to look for explanations for cross-country differences in income and economic growth using new paradigms and ways of processing data. For example, Hausmann, *et al.* (2011) have sought to map the productive capabilities of nations to come up with an Economic Complexity Index (ECI) that captures significantly more information and performs substantially better than other well-known indices of international economic performance.[17] Their visualisation and graphic presentations offer a novel way to interact with the economic data and to explore dynamic changes over time.

This suggests yet another problem that commonly arises in Internet measurement (but is also commonly associated with the measurement of other complex phenomena). In short, because the Internet is a bundle of complementary components that are used in different proportions in different contexts, no single metric is sufficient. Thus, it is common to seek to summarise a mixture of metrics with a composite

index. For example, Atkinson *et al.* (2008) constructed a composite index that combined data on household penetration rates, broadband speeds, and the lowest offered price per Mbps in order to facilitate cross-national broadband rankings. Such composite indices often have great appeal to policymakers interested in rendering the multidimensional phenomena more tractable and understandable. Unfortunately, composite indices can be misleading. The choice of components and their weightings may emphasise some points and obscure others. Analysts differ on the cost/benefit tradeoffs of composite indices, but they are likely to remain an important feature of future Internet policy debates.[18] To help avoid the pitfalls inherent in any such index, it is important that the methods used in constructing any indices be fully disclosed and the underlying data be verifiable.

Democratising the data game

With the proliferation of Internet access, the big data game is a game that increasingly anyone can play. Tools and platforms like the WeFeelFine.org project discussed earlier, Wordle.net,[19] or mash-up tools like Yahoo Pipes[20] provide a range of easy-to-use tools for collecting, processing, and presenting information. New data projects like Google's Measurement Lab[21] are making terabytes of Internet traffic measurement data freely accessible, and with an expanding set of visualisation tools that allow anyone to play with the data. And, not to be outdone, government agencies like the National Telecommunications Information Agency (NTIA) provide free download access to the complete 25 million record database they collected on broadband service availability in the United States.[22] Finally, a number of cloud-service providers like Google, Dropbox, and others are making petabytes of on-line storage freely available where data and analysis tools may be stored and presented. All of this activity is democratising the data analysis game – virtually anyone can collect, analyze, and creatively present large volumes of data on a growing number of policy-relevant issues.

Unfortunately, it is far from clear whether the skill sets of potential users and consumers of all this analysis are keeping pace with the increased availability of tools and data. Whether from ignorance or from calculated misuse, it is increasingly easy to get the data to say almost anything. How persuasive false-data arguments ultimately may be remains to be seen.

The potential problem becomes even scarier when we consider the need to increasingly move toward automated decision-making as we expand our reliance on machine-to-machine (m2m) systems and software-agent-based control systems. As we become further removed from the raw data and an understanding of the methods and models used to process the data, it will become more difficult to detect and correct errors in analysis. Ultimately, this is not a problem created by ICT or the Internet, but a challenge of living in a faster-paced, more integrated, crowded, and complex world.

The Statistical Institution challenge

In the world of data-everywhere, international organisations like the Organisation for Economic Cooperation and Development (OECD), International Telecommunication Union (ITU), the United Nations (UN), and the World Bank, in conjunction with National Statistics Institutions (NSI), have important roles to play as trusted brokers of the public data of record. These organisations play a number of important roles in the collection, analysis, management, and presentation of the data needed by policy-makers for evidence-based decision-making. The traditional model of statistical agencies as data custodians who collect, store, and publish the data is changing to accommodate the growing importance of such agencies role as data curators and communicators. As curators, the statistical agencies need to provide guidance on methods and standards for assessing the merits of alternative data collections, including understanding the appropriate uses (and misuses) of composite indices.

It is beyond the scope and resources of the statistical agencies to collect and maintain all of the data and specialised expertise that will be needed. For example, regulatory agencies, specialist analysts (whether in academia or consultancies), or Internet industry participants may be in a better position to collect, monitor, and interpret data on Internet usage and performance. Statistical agencies may play a role in curating the gathering together of disparate data sources and helping to insure interoperability across data collections.

The traditional cycle of collection, data cleaning and processing, and publication that resulted in much data being made publicly available only after a lag of a year or more is meeting resistance in today's much faster-paced decision-making environments. While it takes time and resources to validate the data, the option of incurring the time and resource costs to proceed in such a linear fashion is no longer a viable option in all cases. The publication of data needs to be more of a two-way communication. The Internet and the electronic accessibility of the data make this a viable option.

In the Internet environment, public accessibility and openness helps facilitate crowd-sourcing data improvements. A more flexible view toward ensuring data quality through continuous improvement, transparency, and open accessibility is more consistent with Internet culture. Of course, through their detailed surveys and censuses of business and consumer usage of ICTs, and through other methods of data collection, the relevant institutions (*e.g.* statistical agencies) will collect far more information than they are likely to be able to make publicly available to all users. Much of the data that is collected under their mandate or is voluntarily provided is made available only under the stipulation that confidentiality and privacy be protected.

One way to address this problem is to have tiered levels of access: for data that cannot be fully made publicly available (the best option when it works), it may be possible to allow more limited access to third-party, independent analysts. Such strategies have been used quite effectively in the past to enable academics to access confidentiality-protected business-establishment survey data. Researchers who work with the raw data are obligated to not disclose results at a level of disaggregation that would allow the raw data to be uncovered. Additionally, there may be technical ways to anonymise the data by recoding or selective sampling to preserve its usefulness for certain types of analysis, while still protecting data privacy. The data security arms race renders the feasibility of such strategies uncertain at this point.

Public-private partnerships in collecting, maintaining, and presenting the data will need to be even more prevalent in the future. Likewise, the Internet data challenge will require more multidisciplinary engagement. For example, to understand data on Internet traffic performance and infrastructure in order to appropriately analyze issues like market power (*e.g.*, for defining Internet interconnection markets or for analyzing behaviour in on-line markets) will require economists to engage with network engineers; similarly, designing appropriate Internet metrics for use in assessing social impacts of Internet usage will require network engineers to engage with social scientists to make sure that the right things are being measured.

Statistical agencies can play an important role in helping to convene interactions across the disparate players. To be effective, policymakers will need to ensure that statistical agencies have adequate resources and expertise to address the difficult analytic challenges of working with large data sets, as well as the Internet technical savvy required to use the Internet effectively as a tool to collect, process and present data electronically.

Statistical agencies also have an important role to play in helping to ensure international harmonisation of data collections. As noted at the beginning of this paper, the first insightful evidence of Internet impacts are likely to be derived from granular, micro data set analyses. These may be completed in isolated markets or countries. Policymakers may wish to know how feasible it is to generalise results

obtained for one market to another. For example, policymakers may wish to know if the estimates from a study of how social networking may be impacting children's educational performance that was conducted in Germany might apply to policy debates in Canada. To allow microanalyses to be leveraged across markets and policy contexts, good standardised and interoperable data will be needed to assess the extent to which the markets may be comparable – and the interoperable data need not be the same data that is most relevant to measuring Internet impacts. When comparability is deemed acceptable, the standardised data may facilitate the appropriate scaling of results. For example, using consistent frameworks for estimating national populations or exchange rates can ease the challenge of scaling/converting results derived from a study conducted in one market for use in another context.

Summing up

Analysis interested in measuring the Internet and its economic and social impact are confronted by a number of challenges and opportunities, some of which have been reviewed in this essay. In addition to all of the normal problems one expects when measuring ICTs, the Internet poses some additional challenges, but also offers some special opportunities.

A first challenge is associated with the continuously evolving Internet that has changed from a service that is used by some to an essential, basic economic infrastructure that is used everywhere by everyone. The focus of data collection needs to move from measuring adoption and availability to measuring usage. Moreover the drive to measure usage will necessitate more granular data. For example, in measuring the economic impacts of the Internet on businesses, the unit of observation is shifting from the firm to the business unit or establishment, to the individual worker, and eventually, to the specific activity being undertaken by the worker, tracked in time and space. Collecting such granular data will necessarily engage policy concerns over privacy and data confidentiality.

Furthermore, significant changes underway in the Internet – to enable mobility, cloud computing resources, social-networking, and sensor networks – imply that tomorrow's critical Internet components are not well-measured by today's Internet metrics. Line counts, fibre miles, megabytes of traffic, or IP addresses are not good proxies for assessing the intensity of the trends identified.

Fortunately, these same trends and the transition of the Internet into basic infrastructure are compelling regulators and value chain participants, including ISPs, to heavily instrument all levels of the Internet to allow fine-grained, real-time control of network traffic. The basic measurement infrastructure to track the Internet is being put in place, even if not by the traditional statistical agencies. The Internet is becoming a powerful platform for data collection, analysis, and presentation for data of all kinds.

The potential of the heavily measurement-instrumented Internet to automate and facilitate data collection, analysis, and presentation processes is illustrated by several projects underway with data as diverse as blog postings expressing feelings to WikiLeaks documents to remote healthcare monitoring technologies. These tools allow for the collection of increasingly larger data sets with significant spatial-temporal granularity.

The transition to “big data” poses a number of additional challenges, including driving the need for cloud-based computing resources and new analytic techniques such as parallel processing. In addition, there is a skills-gap among researchers with respect to how these tools should be used, with the numerical analysis challenges of working with big data, and with the new types of visualisation and data analysis tools that are needed to make sense of the more diverse and complex collections of real-time data that are becoming available.

The openness of the Internet and the accessibility of tools and public data sets are democratising the data game. This holds great potential for expanding public accessibility to evidence and engaging in informed policy debates, but the skills gap problems loom large when un-trained analysts seek to make sense of complex quantitative data. There is an enhanced risk that evidence-based policymaking may suffer either at the hands of those who would abuse the data to strategically mislead public opinion or by growing skepticism about the viability of identifying truths in the data.

In this changing environment, the role of statistical agencies as trusted brokers of public data of record remains important but needs to change. The role of statistical agencies will need to transition from being custodians of the data to curators, from publishers of vetted data to communicators of continuously updated data. In realising this goal, the statistical agencies will need to embrace public-private partnerships and a more fluid and interactive style of data management. And, when it comes to Internet data especially, the data collection/processing/presentation functions will need to be consciously multidisciplinary. Getting good data and metrics for Internet usage for making policy-relevant decisions in a timely manner will require Internet networking engineers and social scientists to actively engage with each other.

The challenge of measuring the Internet and its social and economic impacts is one and the same as the challenge of measuring the overall economy. For those with a taste for tackling interesting questions with novel data and methods, the future is indeed bright.

NOTES

1 In 1987, Robert Solow commented, "we can see the computer age everywhere but in the productivity statistics" (see New York Times, 20 May 1987, p. A1). The measurement of ICT impacts is difficult because both input and output quantities and prices are hard to measure, and because ICTs are General Purpose Technologies (see Bresnahan and Trajtenberg, 1995) that change how goods and services are produced. The measurement difficulties arise for numerous reasons, including the fact that Moore's Law-like productivity improvements in ICTs result in rapid technological progress and economic depreciation; ICT usage is especially intense in the service sectors which are notoriously poorly measured; and the impacts of ICTs take time to be realised.

2 Although both business and consumer Internet usage is of interest, this paper will focus on business usage to explicate the challenges.

3 Internet-enabled telecommuting, business-to-business, and business-to-consumer electronic commerce allow firms organisational structures to be dynamically adjustable.

4 See *www.sandvine.com/*.

5 See *www.arbornetworks.com/*.

6 In addition to policy questions about whether DPI violates subscriber privacy rights (see Ou, 2009,) the different vendors rely on proprietary algorithms and code for traffic measurement and classification that needs to continuously evolve as applications change, motivated in part by a desire to evade traffic limiting controls by operators.

7 See *www.samknows.com/broadband/index.php*.

8 In 2006, the European Union adopted legislation requiring ISPs to retain traffic data, motivated by the needs of law enforcement to protect public safety and national security (see, *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:NOT*).

9 When we seek to measure broadband access to fixed broadband, we naturally focus on household subscriptions, whereas with mobile broadband, we focus on individual subscriptions. In the former case, the service is typically shared with all members of the household; while in the latter, usage is typically (but not necessarily) not shared. As Wallsten (2008) showed, the drop in the United States. broadband ranking from 2002 to 2007 in OECD and other data sources that ranked countries on the population penetration of broadband could be explained on the basis of systematic differences in household sizes. Decline in broadband ranking in the United States was a topic for heated policy debate for several years, prompting a number of papers like Wallsten's that sought to make sense of international broadband comparisons.

10 See *www.wefeelfine.org*. From the mission statement: "Every few minutes, the system searches the world's newly posted blog entries for occurrences of the phrases 'I feel' and 'I am feeling'. When it finds such a phrase, it records the full sentence, up to the period, and identifies the 'feeling' expressed in that sentence (*e.g.* sad, happy, depressed, etc.)…. The result is a database of several million human feelings, increasing by 15 000 – 20 000 new feelings per day. Using a series of playful interfaces, the feelings can be searched and sorted across a number of demographic slices, offering responses to specific questions like: do Europeans feel sad more often than Americans? … At its core, We Feel Fine is an artwork authored by everyone."

[11] See *http://rdn-consulting.com/blog/2010/12/19/the-cardiocam-physiological-monitoring-via-webcam/*. A web-based way to use a webcam to read a person's heart-rate and other medical data remotely.

[12] See "Interactive: Analyze your smile," Forbes.com, March 3, 2011 (available at: *www.forbes.com/2011/02/28/detect-smile-webcam-affectiva-mit-media-lab.html*). A Web-based application to use your computer's camera to track your emotions over time to identify what portions of an advertisement you found amusing.

[13] See "Big data, big problems: the trouble with storage overload," GIZMODO, March 17, 2010 (available at: *http://gizmodo.com/5495601/big-data-big-problems-the-trouble-with-storage-overload*).

[14] Examples of new software tools include MapReduce and Hadoop. A recent article reported that MapReduce was successful in sorting a petabyte file of 100-byte records on a system of 8 000 computers in 33 minutes compared to the six hours it took to accomplish the same task on a cluster of 4 000 machines in 2008 (see "Sorting Petabytes with MapReduce – The Next Episode," September 2011, available at: *http://googleresearch.blogspot.com/2011/09/sorting-petabytes-with-mapreduce-next.html*). While the tools keep getting faster, the data sets are growing larger.

[15] According to Chris Anderson, "Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." (see Anderson, 2008.)

[16] For a discussion of both good and bad practices when trying to present quantitative information visually, see Tufte and Howard (1983).

[17] Hausmann, *et al.* (2011) find their ECI is significantly better at explaining the relative economic performance of nations during the 2002-2007 period than the World Bank's Worldwide Governance Index (WGIs, see *http://info.worldbank.org/governance/wgi/index.asp*) or the World Economic Forum's Global Competitiveness Index (GCI, see *www.weforum.org/issues/global-competitiveness*).

[18] For examples of popular Internet indices, consider the following: EIU, 2010; Waverman, 2011; BGG, 2010; ITU, 2011; McKinsey Global Institute, 2011; and most-recently W3C (2011). All of these indices seek to integrate multiple supply and demand side metrics of Internet performance that may prove helpful in explaining relative performance differences across countries.

[19] See *www.wordle.net/* is a "toy for generating 'word clouds' from user-provided text" that allows users to quickly generate graphic images that highlight prevalent words in the text.

[20] See *http://pipes.yahoo.com/pipes/*, a mash up tool from Yahoo that allows you to combine RSS feeds, reorganise, filter, and sort Web-based information, and integrate other Web 2.0 interactive tools for user-generated analysis tools on the fly.

[21] *www.measurementlab.net/*.

[22] See *www.broadbandmap.gov/data-download*. The NTIA data includes 25 million records listing the availability by service provider and speed-tier by Census Block, allowing one to discern which service providers are offering what level of service at a very fine level of granularity. The NTIA includes a number of APIs for analyzing and presenting the data, as well as linking it to other data sources, including speed-testing data from Google's Measurement Lab.

REFERENCES

Anderson, C. (2008) "The end of theory: the data deluge makes the scientific method obsolete," *Wired Magazine*, June 23, available at: *www.wired.com/science/discoveries/magazine/16-07/pb_theory.*

Atkinson, R., D. Correa, and J. Hedlund (2008), "Explaining International Broadband Leadership," Information Technology & Innovation Foundation (ITIF), Washington, DC, May 1, 2008, available at: *http://archive.itif.org/index.php?id=142*.

Bauer, S., D. Clark, and W. Lehr (2010), "Understanding Broadband Speed Measurements," MITAS Working Paper, June 2010, available at: *http://mitas.csail.mit.edu/papers/Bauer_Clark_Lehr_Broadband_Speed_Measurements.pdf.*

Boston Consulting Group (BCG) (2010), "The Connected Kingdom: How the Internet is Transforming the U.K. Economy," report prepared by Boston Consulting Group, available at: *www.bcg.com/documents/file62983.pdf*).

Bresnahan, T. and M. Trajtenberg (1995) "General Purpose Technologies: Engines of Growth," *Journal of Econometrics*, vol. 65, pp. 83-108

Brynjolfsson, E. and L. Hitt (1998), "Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending," *Management Science*, April, reprinted in L. Lillcocks and S. Lester (eds.), *Beyond the IT Productivity Paradox: Assessment Issues*, McGraw Hill, Maidenhead, 1998.

Economist Intelligence Unit (EIU) (2010), "Digital Economy Rankings 2010: Beyond e-readiness," EIU, London, available at: *http://graphics.eiu.com/upload/EIU_Digital_economy_rankings_2010_FINAL_WEB.pdf*.

Hausmann, R., C. Hidalgo, S. Bustos, M. Coscia, S. Chung, J. Jimenez, A. Simoes, and M. Yildirim (2011), "The Atlas of Economic Complexity: Mapping Paths To Prosperity", MIT, September 2011, available at: *http://atlas.media.mit.edu*/.

International Telecommunications Union (ITU), (2011), Measuring the Information Society: the ICT Development Index (IDI), ITU, Geneva, available at: *www.itu.int/ITU-D/ict/publications/idi/2011/Material/MIS_2011_without_annex_5.pdf*.

Jorgenson, D. (2001), "Information Technology and the U.S. Economy," *American Economic Review*, vol. 91, no. 1, March, pp. 1-33.

Judd, K. (1998), *Numerical Methods in Economics*, MIT Press: Cambridge, MA

Lehr, W. and F. Lichtenberg (1999), "Information Technology and Its Impact on Productivity: Firm-level Evidence from Government and Private Data Sources, 1977-1993", *Canadian Journal of Economics*, vol. 32, no. 2, April, pp. 335-362

Lehr, W., S. Bauer, M. Heikkinen, and D. Clark (2011), "Assessing Broadband Reliability: Measurement and Policy Challenges," 39th Research Conference on Communications, Information and Internet Policy (*www.tprcweb.com*), Alexandria, Virginia, September, *http://people.csail.mit.edu/wlehr/Lehr-Papers_files/Lehr%20et%20al%20TPRC2011%20Assessing%20Broadband%20Reliability.pdf*.

McKinsey Global Institute (2011), "Internet Matters: the Net's Sweeping Impact on Growth, Jobs and Prosperity," McKinsey Global Institute, May, available at: *www.mckinsey.com/mgi/publications/internet_matters/pdfs/MGI_internet_matters_full_report.pdf*.

Oliner, D. and D. E. Sichel (2000), "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story ?", Federal Reserve Board Finance and Economics, Discussion Series 2000/20, March.

Ou, G. (2009), "Understanding Deep Packet Inspection (DPI) Technology," a White Paper from Digital Society, October 2009, available at: *www.digitalsociety.org/files/gou/DPI-Final-10-23-09.pdf*.

Schactman, N. (2010), "Open Source Tools Turn WikiLeaks into Illustrated Afghan Meltdown (Updated)," Wired, August 10, 2010, available at: *www.wired.com/dangerroom/2010/08/open-source-wikileaked-docs-illustrated-afghan-meltdown/*.

Sen, S., J. Yoon, J. Hare, J. Ormont, and S. Banerjee (2011), "Can you hear me now? A case for client-assisted approach to monitoring wide-area wireless networks," Internet Measurement Conference (IMC '11), Berlin, Germany, November 2-3, available at: *http://pages.cs.wisc.edu/~suman/pubs/wiscape.pdf*.

Tesfatsion, L., and K. Judd (2006), *Handbook of Computational Economics: Agent-based Computational Economics*, North-Holland: New York.

Tufte, E. and G. Howard (1983), *The Visual Display of Quantitative Information*, Graphics Press: Cheshire, Connecticut.

World Wide Web Foundation (W3C) (2011), "The World Wide Web Index", W3C, available at: *www.webfoundation.org/projects/the-web-index*.

Wallsten, S. (2008), "Understanding International Broadband Comparisons", Technology Policy Institute Working Paper, Washington, DC, May 2008, available at: *www.techpolicyinstitute.org/files/wallsten_international_broadband_comparisons.pdf*.

Waverman, L. (2011), Connectivity Scorecard 2011, a study prepared for Nokia-Siemens Networks, available at: *www.connectivityscorecard.org*.